一种多粒度空间的快速构建方法

赵 凡[1),3)] 张清华[1),2),3),4)] 吴成英[1),2)] 谢 秦[1),2)] 王国胤[1),2)]

[1)](重庆邮电大学计算智能重庆市重点实验室 重庆 400065)

[2)](旅游多源数据感知与决策技术文化和旅游部重点实验室 重庆 400065)

[3)](重庆邮电大学大数据智能计算重点实验室 重庆 400065)

[4)](重庆邮电大学网络空间大数据智能安全教育部重点实验室 重庆 400065)

摘 要 粒计算是模拟人脑多粒度认知模式处理复杂问题的一种方法.模糊商空间理论作为粒计算的一种典型模型,将复杂问题渐进式粒化成为分层递阶的多粒度空间,从而实现层次化的求解.然而,面对海量高维数据,现有模糊商空间模型通过模糊相似关系构建多粒度空间的效率将大幅降低.一方面,模糊相似关系需要计算数据空间中任意两个对象之间的相似性,不利于处理体量大的数据集;另一方面,模糊相似关系包含大量冗余信息,导致后续步骤中存在大量的冗余计算.因此,本文基于2近邻模糊关系,提出了多粒度空间的快速构建方法,在保证面向下游分类任务时性能不下降的前提下,极大地提升了多粒度空间构建效率.首先,基于 k 近邻算法提出 k 近邻模糊关系,并分析证明其关键性质;然后,面向多粒度空间构建任务,对 k 近邻模糊关系进行参数分析,从理论上证明 k 取2时即可包含数据空间中全部有效信息;随后,定义了最近邻和次近邻两阶段的有效位置数,提出了模糊相似关系有效值和有效位置提取算法,多粒度空间构建效率提升了75%左右.最后,通过在9个UCI数据集、3个UKB数据集、3个图像数据集和3个文本数据集上的相关实验,验证了该算法构建多粒度空间的高效性、正确性以及面向下游分类任务的有效性、稳定性和显著性.

关键词 粒计算;多粒度空间;k 近邻;模糊关系;模糊商空间

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2024.02141

An Efficient Approach for Constructing the Multi-Granularition Spaces

ZHAO Fan[1),3)] ZHANG Qing-Hua[1),2),3),4)] WU Cheng-Ying[1),2)]
XIE Qin[1),2)] WANG Guo-Yin[1),2)]

[1)](*Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing* 400065)

[2)](*Key Laboratory of Tourism Multisource Data Perception and Decision, Ministry of Culture and Tourism, Chongqing* 400065)

[3)](*Key Laboratory of Big Data Intelligent Computing, Chongqing University of Posts and Telecommunications, Chongqing* 400065)

[4)](*Key Laboratory of Cyberspace Big Data Intelligent Security, Ministry of Education, Chongqing University of Posts and Telecommunications, Chongqing* 400065)

Abstract Granular computing is a state-of-the-art methodology that simulates the multi-granularition cognitive pattern of the human brain to deal with complex problems. As a typical description of granular computing, fuzzy quotient space theory focuses on gradually granulating complex problems into the hierarchical multi-granularition spaces, thereby implementing hierarchical solution of the complex problems. However, when dealing with massive high-dimensional data, the efficiency of constructing multi-granularition spaces through the fuzzy similarity relations in the

收稿日期:2023-08-29;在线发布日期:2024-05-22.本课题得到国家自然科学基金(No. 62276038,No. 62221005)和重庆市自然科学基金(No. cstc2019jcyj-cxttX0002).**赵 凡**,博士研究生,主要研究领域为不确定性信息处理与度量.E-mail:837062256@qq.com.**张清华**(通信作者),博士,博士生导师,教授,中国计算机学会(CCF)会员,主要研究领域为粒计算、数据挖掘和不确定性信息处理.E-mail:zhangqh@cqupt.edu.cn.**吴成英**,博士研究生,主要研究领域为粒计算与监督学习.**谢 秦**,博士研究生,主要研究领域为知识发现与粒计算.**王国胤**,博士,教授,博士生导师,中国计算机学会(CCF)会士,长江学者,主要研究领域为粒计算、数据挖掘和神经网络.

existing fuzzy quotient space methods reduces significantly. On one hand, the fuzzy similarity relation is obtained through calculating the similarity among all objects, which is not conducive to processing large datasets; On the other hand, the fuzzy similarity relation contains a large amount of redundant information, which leads to a large number of redundant computation in the subsequent steps. Therefore, based on the 2-nearest neighbor fuzzy relation, an efficient construction approach for constructing multi-granularition spaces is proposed, which greatly improves the efficiency on the premise of ensuring the performance when facing downstream classification tasks. First, based on the k -nearest neighbor algorithm, a k -nearest neighbor fuzzy relation is proposed, and its key properties are discussed and proven. Second, for the multi-granularition spaces construction task, parameter analysis is performed on the k -nearest neighbor fuzzy relation, theoretically proving that when k is taken as k , all effective information in the data space could be included. Then, the number of effective positions in the nearest neighbor and second nearest neighbor phases are defined. And the algorithm for extracting effective values and effective positions of the fuzzy similarity relation is proposed, the efficiency of constructing multi-granularition spaces is improved by about 75%. Finally, relevant experiments are conducted on 9 UCI datasets, 3 UKB datasets, 3 image datasets and 3 text datasets to validate the efficiency of multi-granularition spaces construction approach. By comparing and analyzing with the existing classifiers, the effectiveness, stability, and saliency of the proposed approach for classification tasks are demonstrated. In summary, a k -nearest neighbor fuzzy relation that only contains sparse effective information is constructed. On the basis, an efficient construction approach for constructing multi-granularition spaces based on 2-nearest neighbor fuzzy relation is proposed, which greatly reduces time complexity while ensuring classification performance.

Keywords granular computing; multi-granularition spaces; k -nearest neighbors; fuzzy relation; fuzzy quotient space

1 引 言

随着世界进入智能化时代,源源不断的数据呈井喷式增长,海量数据给人类提供方便的同时也带来了一系列棘手的问题.因此,如何准确地从海量数据中获取隐含的有价值的信息是知识发现算法的研究热点.随着算力的迅速发展和模型的不断优化,知识发现算法已取得众多重大突破.然而,数据的海量特性与其所蕴含知识的稀疏性之间的相持、数据的强动态变化与知识的弱演化能力之间的对立[1]、数据的价值随时间骤减的高时效性等仍然为知识发现算法带来极大的挑战.因此,如何有效且快速地运用数据资源实时提取知识、挖掘更高的数据价值是知识发现领域亟需解决的科学问题.

粒计算(Granular computing,GrC)是人工智能领域中一种以人为本的先进方法论,旨在通过模拟人脑认知机制来处理复杂数据[2-4].作为一种知识发现的新兴技术,GrC 能够从不准确、不完整、不确定的海量数据中挖掘出蕴含的多层次知识还原现实世界以应对复杂问题[5].早在 1979 年,美国工程院士 Zadeh 教授在论文“Fuzzy sets and information granularity”中指出众多领域都存在信息粒的现象[6],随后在 1998 年,Lin 教授[7]正式提出了 GrC 的概念.自提出以来,GrC 引起了学术界的广泛关注,已成为不确定信息处理、大尺度计算、云计算等领域的重要工具[8-10].GrC 通过模拟人脑进行观察感知、尺度度量、概念内化和决策推理时的“大范围优先”和“渐进式”等特性,将数据空间划分为不同层级的颗粒来构造多粒度空间,实现对问题由浅入深的认知和渐进式求解[11].鉴于此,通过渐进式粒化构建多粒度空间和在多粒度空间中进行粒的计算成为粒计算的两个重要方向[12,13].因此,面向复杂数据构建多粒度空间成为 GrC 领域的研究热点.

1992 年清华大学张钹院士提出的商空间理论[14],该理论通过模拟人类的多粒度、多层次思维模式构建还原数据拓扑结构的商空间,为多粒度空间提供了较为通用的公理化定义和构造方法,是经

典的 GrC 模型之一. 随后,为了处理连续数据,张钹院士将模糊关系引入商空间理论提出模糊商空间理论[15],提高了模型的鲁棒性. 模糊商空间通过建立基于模糊等价关系的多层次拓扑结构来模拟人类在不同粒度下解决复杂问题的能力,不仅能够挖掘数据中的隐含知识以建立多粒度空间,而且能够刻画复杂问题在不同粒度之间的转换和相互依赖关系. 一方面,模糊商空间理论可将不确定的概念细化到多粒度空间中,称为层次商空间结构(Hierarchical Quotient Space Structure,HQSS)[16]. 面对复杂数据,HQSS 通常将不确定性概念以自上而下的方式在不同的知识空间中分解,得到局部解后以自下而上的方式进行整合得到原始问题的全局解[17-18],既模拟了人脑"大范围优先"的认知模式,又运用了计算机由细到粗的计算模式[1]. 另一方面,模糊商空间理论揭示了模糊集的多粒度特性,为模糊集的结构化定义提供了理论基础. 众所周知,在模糊集的应用中,选择不同的隶属度函数可能导致数据对不确定性概念的隶属度存在差异,为后续的数据处理带来极大的不确定性,因此确定合适的隶属度函数是模糊集应用中的关键问题. 而模糊等价关系对应的 HQSS 为隶属度函数的构造提供了一个新的视角,若不确定性概念对应数据的隶属度之间的偏序关系相同或 HQSS 相同,则对应的模糊集具有相同或相似的特征[19]. 因此,HQSS 不仅为模糊数据处理提供了一种粒度处理范式,也为模糊集理论中研究隶属函数的鲁棒性提供了强有力的理论基础.

自提出以来,模糊商空间理论通过构建多粒度空间求解复杂问题的方法论引起了学者的广泛关注. 在构建多粒度空间方面,Miao 等人[20]引入知识粒度构造模糊等价关系,进而优化 HQSS 的构造;Qian 等人[21]从粗糙近似的角度出发,引入多粒度粗糙集模型从上下近似的视角构造 HQSS,并将该模型用于处理不完备数据[22]. Lin 等人[23]将证据理论与多粒度粗糙集相结合,从统计学的角度构建 HQSS. Zhao 等人[24]将模糊商空间中的模糊等价关系扩展为加权等价关系和容差关系,提出了面向层次坐标数据的 HQSS 构建方法,为高效搜索大型网络的最优路径提供知识基础. Huang 等人[25]引入直觉模糊粗糙集来处理特殊数据,并建立直觉模糊多粒度粗糙集模型来构造 HQSS. Yang 等人[26]引入测试代价,提出了双代价敏感的多粒度空间构建模型. Wu 和 Leung[27]引入多尺度决策系统来构建 HQSS,并将其用于知识获取. Li 和 Hu[28]提出了不同属性对应不同数量尺度的广义多尺度决策系统来构建 HQSS,进一步提出网格模型来搜索 HQSS 的所有最优尺度组合;Zhang 等人[29]将三支决策与 Hasse 图相结合应用到 HQSS 的最优尺度组合选择中,极大地提升了知识发现的效率;Zhang 等人[30]将属性代表的概念引入多粒度空间构建中,并将其应用到集成分类中;Xia 等人[31]引入 *k-means* 算法以改进邻域粗糙集的构建机制,提出了粒球邻域粗糙集模型实现不同粒度上的邻域构建;Wu 等人[32]引入密度峰值聚类优化多粒度邻域的构建,提出了超区间粒化方法构建多粒度邻域空间,并设计了相应的分类算法;Zhang 等人[33]通过分析并证明模糊相似关系有效值和有效位置的存在性,提出了提取有效信息的快速算法,极大地提升了 HQSS 的构建效率. 在多粒度空间的度量方面,Liang 等人[34]通过度量知识粒度的不确定性描述 HQSS;Qian 等人[35]通过研究 HQSS 的知识结构,提出了多粒度空间中粒层之间的知识距离度量. Yao 等人[36]通过度量所有划分的熵来度量 HQSS 的不确定性;Zhang 等人[37]提出了不同多粒度空间之间的距离;基于上述研究,Zhang 和 Wang 等人[38]引入了一个信息熵序列来表征多粒度空间中多个粒层的不确定性;Yang 等人[39]从概率粗糙集模型三个区域的角度度量分析了多粒度空间的不确定性;Zhang 等人[40]提出了任意两个多粒度空间之间的分类同构的概念,并讨论了分类同构的充要条件. Yang 等人[41]提出了一种基于推土机距离(Earth mover's distance,EMD)来度量多粒度空间中划分之间的关系;Zhao 等人[42]结合纯度和复杂度提出了多粒度空间粒层的质量度量方法,并基于此进行最优粒度选择.

随着理论研究的不断深入,通过模糊商空间理论构建多粒度空间的方法也得到了广泛的应用. Pedrycz[43]将多粒度空间层次递阶的信息提取后引入到层次模糊 C 均值聚类算法中. Tsekouras 等人[44]分析了 HQSS 和模糊聚类之间的关系,并提出了一种用于模糊建模的多粒度模糊聚类方法. Zhang 等人[45]提出了一种基于 HQSS 的层次模糊决策方法. Tang 等人[46]分析了多粒度空间中随着粒度变化粒层之间的模糊概率关系,并将其应用于层次聚类. Cui 等人[47]通过改进的命令滤波反演法将 HQSS 引入到输入饱和的 mimo 非线性系统中,实时处理信号数据. 综上所述,HQSS 已被应用于包括模糊控制[48-50]、智慧医疗[51]、图像处理[52-53]、模糊逻辑[54-56]等众多领域.

然而,现有的构建多粒度空间的几类方法分别存在不同程度的缺陷:(1) 通过多粒度粗糙集和多尺度粗糙集模型构建多粒度空间的方法只能处理离散型数据,面对连续型数据时需要首先进行离散化,这一过程存在极大的不确定性,同时影响下游分类任务的性能;(2) 通过邻域粗糙集模型构建多粒度空间的方法虽然能够处理连续型数据,但是其中的规则存在大量的包含和重叠关系,必须进行时间复杂度较高的规则约简才能用于下游任务;(3) 通过粒球计算构建多粒度空间的方法,其中存在大量的只包含一个对象的规则,无法形成有效的论域覆盖,这将影响下游分类任务的性能,同时该方法需要人为给定参数,参数的不同将很大程度上影响分类效果;(4) 通过经典模糊商空间模型和基于有效值的模糊商空间模型构建多粒度空间的方法既能够处理连续性数据,同时能够形成有效的论域覆盖,且不存在超参数.在前序工作[42]中,我们已将其成功应用于下游分类任务,提出了不含超参数、不需要规则约简且能够有效覆盖论域的邻域覆盖分类器,并验证了其分类性能优于现有 GrC 分类器和经典机器学习分类器,然而,基于模糊商空间构建多粒度空间的方法在效率上仍然有待提升,具体来看,现有模糊商空间方法的步骤和对应时间复杂度见表 1,存在如下问题:

表 1 现有模糊商空间模型步骤及其时间复杂度

经典模糊空间模型		基于有效值的模糊商空间模型	
步骤	时间复杂度	步骤	时间复杂度
输入:模糊决策信息系统		输入:模糊决策信息系统	
步骤 1:模糊相似关系	$O(n^2)$	步骤 1:模糊相似关系	$O(n^2)$
步骤 2:模糊等价关系	平方法:$O(n^5)$ Warshall 算法:$O(n^3)$	步骤 2:模糊等价关系	$O(mn^2)$
步骤 3:截关系	$O(pn^2)$	步骤 3:截关系	$O(pn^2)$
步骤 4:HQSS	$O(pn^2)$	步骤 4:HQSS	$O(pn^2)$

(1) 步骤 1(构建模糊相似关系):现有模糊商空间方法均需计算数据空间中所有对象两两之间的相似性,构建出 $n \times n$ 的模糊相似矩阵.一方面,该过程对数据空间的对象数和属性数极其敏感,当面对海量数据时,该过程将耗费大量的时间;另一方面,据文献[33]中的理论分析:模糊相似关系共有 n^2 个信息,但其中至多包含 $n-1$ 个有效信息,这说明该过程耗费了大量无效时间;

(2) 步骤 2:经典模糊商空间模型通过平方法($O(n^5)$)、Warshall($O(n^3)$)算法等求模糊相似关系的传递闭包来得到模糊等价关系,这些求传递闭包的算法的时间复杂度均较高;基于有效值的模糊商空间模型通过理论分析证明了模糊相似关系中有效信息的数目和位置,提出了快速提取模糊相似关系有效值的算法,避免了进行求传递闭包的复杂运算的同时,达到了与模糊等价关系相同的粒化效果,已将该步骤的时间复杂度降低至 $O(n^2)$.然而,只加速该步骤无法缓解其他步骤的效率;

(3) 步骤 3(构建关系)和步骤 4(构建多粒度空间 HQSS):经典模糊商空间模型需要依次用 k 个阈值处理模糊等价关系中的 n^2 个值,基于有效值的模糊商空间模型需要依次用 k 个阈值处理模糊相似关系中的 n^2 个值,得到 k 个截关系;进而需要处理 k 个截关系,对论域进行 k 次划分,得到 k 层的多粒度空间.这两个步骤由于模糊相似关系和模糊等价关系的有效信息稀疏性,均需要执行大量的无效循环.

基于以上分析,通过模糊商空间理论构建多粒度空间的研究迫切需要从步骤 1 进行加速,提出更加高效的多粒度空间构建模型.据文献[33]的理论证明,模糊相似关系的有效信息为除主对角线元素外的行(列)最大值和其余值的最大值.因此,为了解决上述问题,本文首先引入 k 近邻算法,构建只包含稀疏有效信息的 k 近邻模糊关系,在此基础上提出基于 2 近邻模糊关系的多粒度空间快速构建方法,实现了在保证效果的前提下大幅降低时间复杂度.具体地,本文的贡献点如下:

(1)理论分析(2 近邻模糊关系包含模糊相似关系全部有效信息).引入 k 近邻算法,提出了 k 近邻模糊关系的概念,并分析其自反性、对称性和传递性;进一步研究并证明了 k 近邻模糊关系随 k 的变化呈现的规律;针对 k 的取值,通过分析模糊相似关系有效值的提取算法,从理论上证明了 2 近邻模糊关系包含了模糊相似关系全部的有效值,并等价于模糊相似关系对应的模糊等价关系,且能够构建出与现有模糊商空间模型完全相同的多粒度空间.鉴于此,将构建 2 近邻模糊关系作为模糊商空间模型构建多粒度空间的步骤 1,相比现有方法,既提升了步骤 1 的效率,又为后续步骤的加速提供了理论基础;

(2)模型构造(基于 2 近邻模糊关系两阶段提取有效信息).基于 2 近邻模糊关系,提出了论域的最近邻序列和次近邻序列,基于相互最近邻数设计最近邻阶段的有效位置数计算方法,基于统计学的频数概念设计次近邻阶段的有效位置数计算方法,并在此基础上提出了两阶段的有效值有效位置提取算法,以加速模糊商空间模型构建多粒度空间的步骤 2;

(3)算法设计(提出多粒度空间快速构建算法).基于上述理论分析,提出基于 2 近邻模糊关系的多粒度空间快速构建方法,相比现有的模糊商空间模型,所有步骤的时间复杂度均得到了降低.最后通过对比实验,验证本文方法构建多粒度空间的效率和正确性.进一步将该算法应用到分类任务中,在 4 个分类指标下验证了本文方法面向分类任务的有效性、稳定性和显著性.因此,相比现有模糊商空间模型,本文方法提升了多粒度空间的构建效率,并能够构建出与现有模糊商空间模型完全一致的多粒度空间,保证了与现有模糊商空间模型相同的分类能力:平均分类性能优于其他粒计算分类器和机器学习分类器.

本文第二节介绍与本文相关的背景知识,主要涉及模糊商空间理论以及 k 近邻的基础定义;第三节介绍本文提出的 k 近邻模糊关系、两阶段的有效位置提取方法以及多粒度空间快速构建方法;第四节主要通过实验对比验证所提算法的效率和有效性.

2　背景知识

为了便于理解,本节首先归纳了本文使用的概念和符号,如表 2 所示.然后,回顾了本文模型涉及的基础知识,主要包括模糊决策信息系统、模糊商空间和 k 近邻等.

为了便于描述和研究,本文分类任务中的数据空间均由模糊决策信息系统表示.

定义 1.模糊决策信息系统[47].给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,其中 U 为一个非空有限集合,称为论域;$At=C\cup D$ 代表属性集,其子集 C 代表条件属性集,子集 D 代表决策属性集;V 代表属性值集,且任一属性的属性值均为模糊集;$f:U\times C\rightarrow V$ 代表描述论域中对象属性值的映射.

表 2　概念符号

符号	概念
$FIDS$	模糊决策信息系统
n	$FIDS$ 中的对象数
$\overline{R}$	模糊二元关系,简称模糊关系
$\widehat{R}$	模糊相似关系
R	模糊等价关系
$x\wedge y$	$x\wedge y=\min(x,y)$
$x\vee y$	$x\vee y=\max(x,y)$
$\lvert X\rvert$	集合 X 的势
R_{knn}	k 近邻模糊关系
$R_{knn,\lambda}$	k 近邻模糊关系的截关系
$H(\cdot)$	$\cdot$ 的值域
m	次近邻阶段的有效值个数
p	有效值个数
x^k	距离对象 x 第 k 近的对象
$Eff(\cdot)$	$\cdot$ 的有效值集合
$Eff_p(\cdot)$	$\cdot$ 的有效位置集合
$1st(U)$	最近邻序列
$2nd(U)$	次近邻序列
$\lvert Eff_{p1}(R_{2nn})\rvert$	最近邻阶段的有效位置数
$\lvert Eff_{p2}(R_{2nn},i)\rvert$	次近邻阶段第 i 次的有效位置数

2.1　模糊商空间

模糊商空间是一种典型的粒计算模型,能够层次递阶地挖掘模糊数据中的结构化信息,进而构建多粒度空间.

2.1.1　经典模糊商空间模型

给定一个 $FDIS=<U,At,V,f>$,经典模糊商空间模型构建多粒度空间的步骤如图 1 所示.

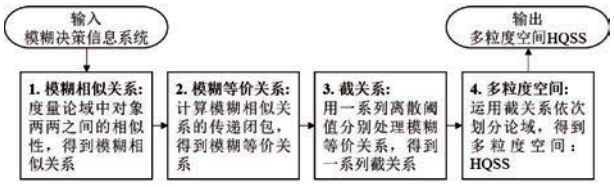


图 1　经典模糊商空间模型构建多粒度空间

由于篇幅限制,更详细的构建过程见文献[37].相关定义如下:

定义 2.模糊相似关系[48,50].给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,令 $\overline{R}$ 为 U 上的模糊二元关系.若 $\overline{R}$ 满足以下条件:

(1) 对于 $\forall x\in U,\overline{R}(x,x)=1$;

(2) 对于 $\forall x,y\in U,\overline{R}(x,y)=\overline{R}(y,x)$,

则 $\overline{R}$ 为模糊相似关系,记作 $\widehat{R}$.

定义 3.模糊等价关系[48,50].给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,令 $\overline{R}$ 为 U 上的模糊二元关系.若 $\overline{R}$ 满足以下条件:

(1) 对于 $\forall x\in U,\overline{R}(x,x)=1$;

(2) 对于 $\forall x,y \in U, \overline{R}(x,y)=\overline{R}(y,x)$;

(3) 对于 $\forall x,y,z \in U, \overline{R}(x,z) \geqslant \sup_{y \in U} \min(\overline{R}(x,y), \overline{R}(y,z))$,

则 $\overline{R}$ 为模糊等价关系,记作 R.

定义 4. 截关系[50]. 给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,令 R 为论域 U 上的模糊等价关系,则 $R_\lambda=\{(x,y) \mid R(x,y) \geqslant \lambda\}(0 \leqslant \lambda \leqslant 1)$ 为 R 上的截关系.

由截关系 R_λ 得到的划分称为知识空间,记作 $\pi_{R_\lambda}(U)=U/R_\lambda$.

定义 5. 分层递阶的商空间结构[37]. 给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,令 R 为论域 U 上的模糊等价关系,则集合 $\pi_R(U)=\{U/R_\lambda \mid \lambda \in H(R)\}$ 称为 R 的分层递阶的商空间结构,简记作 HQSS.

值得注意的是,分层递阶的商空间结构即为一种典型的多粒度空间.

2.1.2 基于有效值的模糊商空间模型

给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,基于有效值的模糊商空间模型构建多粒度空间的步骤如图 2 所示.

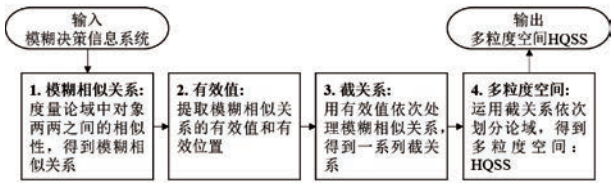


图 2 基于有效值的模糊商空间模型构建多粒度空间

由于篇幅限制,更详细的构建过程见文献[33]. 相关定义如下:

定义 6. 有效值和有效位置[33]. 给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,令 $U=\{x_1,x_2,\ldots,x_n\}$,$\widetilde{R}$ 为 U 上的模糊相似关系且 R 为 U 上对应的模糊等价关系. 对于 $\forall \widetilde{R}(x_i,x_j)(i \neq j)$,若 $\exists h \in H(R)$ 满足以下条件:

$$\widetilde{R}(x_i,x_j)=h \tag{1}$$

则 $\widetilde{R}(x_i,x_j)$ 称为 $\widetilde{R}$ 的有效值,且 $[i,j]$ 称为 $\widetilde{R}$ 的有效位置. $\widetilde{R}$ 的有效值集和有效位置集分别记作 $Eff(\widetilde{R})$ 和 $Eff_p(\widetilde{R})$.

2.2 k 近邻

对象与其近邻之间的关系常常作为许多知识发现算法的基础,可以有效提高算法的效果[57],下面将介绍本文所提出的算法中用到的几种近邻关系.

定义 7. k 近邻[58]. 给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,对于 $\forall x \in U$,则

$$knn(x)=\{\forall y \in U \mid d(x,y) \leqslant d(x,x^k)\} \tag{2}$$

为 x 的 k 近邻,其中 $d(x,y)$ 表示 x 和 y 之间的距离和 x^k 表示距离对象 x 第 k 近的对象.

定义 8. 相互 k 近邻[59]. 给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,对于 $\forall x \in U$,则

$$mknn(x)=\{\forall y \in U \mid x \in knn(y) \wedge y \in knn(x)\} \tag{3}$$

为 x 的相互 k 近邻,其中 $d(x,y)$ 表示 x 和 y 之间的距离.

3 基于 2 近邻模糊关系的多粒度空间快速构建方法

为了实现基于模糊商空间模型的多粒度空间快速构建,本文 3.1 节定义了 k 近邻模糊关系,并面向多粒度空间构建任务对参数 k 进行了取值分析,进而定义了最近邻序列和次近邻序列;3.2 节基于 2 近邻模糊关系、最近邻序列和次近邻序列,设计了两阶段的有效值和有效位置提取算法;3.3 节提出了多粒度空间的快速构建算法,并与现有算法的时间复杂度进行对比分析.

3.1 基于 2 近邻模糊关系的多粒度空间快速构建方法

由本文 2.1 节中图 1 和图 2 可知,现有模糊商空间模型构建多粒度空间的方法的步骤均是首先计算论域中两两对象之间的相似度,得到模糊相似关系. 然后通过不同的方法提取其中的有效信息,得到一系列不同的阈值,最后通过使用阈值逐一对论域进行划分,来构建多粒度空间. 然而,一方面,步骤 1 计算论域内两两对象之间相似度的时间复杂度为 $O(n^2)$,由于相似性度量的计算特性,该步骤对论域的对象数和属性数均极其敏感,因此当面对海量高维数据时,步骤 1 将耗费大量的时间;另一方面,据文献[33]中的理论分析:模糊相似关系包含的 $n \times n$ 个信息中,至多有 $n-1$ 个有效信息,这说明步骤 1 耗费的大量时间是无效的. 因此,基于模糊商空间模型设计多粒度空间快速构建方法应当首先着眼于步骤 1 的加速.

文献[33]中关于模糊相似关系有效值和有效位置的理论证明和算法的设计,如图 3 所示.

由图 3 可知,提取模糊相似关系中的有效值大

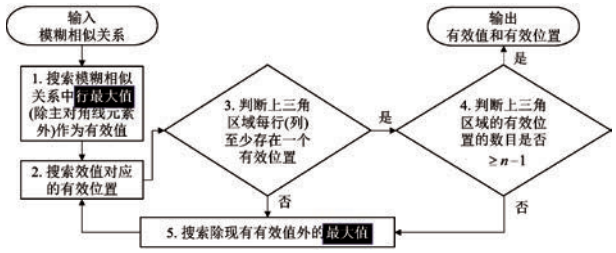


图 3　模糊相似关系的有效值和有效位置提取[33]

致分为两个阶段:首先提取出主对角线外的行(列)最大值,然后对剩余元素依次提取最大值.换言之,模糊相似关系的无效计算大多耗费在所有行的较小值和全局较小值的计算上.而行(列)最大值代表的是与所有对象与其最接近的对象之间的相似度,剩余元素的全局最大值即为相似度较高两个对象之间的相似度.基于上述分析,3.1.1 节将基于所有对象的 k 近邻信息,构建"富含"有效信息的模糊关系,称为 k 近邻模糊关系,并研究其相关性质;3.1.2 节将面向多粒度空间快速构建这一任务,对 k 近邻模糊关系中的 k 进行参数分析,证明 2 近邻模糊关系即可包含原模糊相似关系中全部的有效信息;进一步地,定义了最近邻序列和次近邻序列.既加速了模糊商空间模型构建多粒度空间时的步骤 1,又为后续步骤的加速提供了良好的基础.

3.1.1　k 近邻模糊关系及其性质

在本节中,在模糊决策信息系统上定义了 k 近邻模糊关系,研究并证明了其相关性质,为基于模糊商空间模型快速构建多粒度空间提供理论基础.

定义 9. k 近邻模糊关系.给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,设 $|U|=n$.规定 $U\times U$ 上的模糊子集为 $\overline{R}_{knn}$,对于 $x_i,x_j\in U$,都有

$$\overline{R}_{knn}=\begin{cases}\widetilde{R}(x_i,x_j), & x_j\in knn(x_i)\\ 1, & i=j\\ 0, & \text{otherwise}\end{cases}\tag{4}$$

其中,k 为自然数且 $k<n$,$knn(x_i)$ 代表对象 x_i 的 k 近邻集合(不含自身),$\widetilde{R}(x_i,x_j)$ 代表对象 x_i 与 x_j 之间的相似度.则 $\overline{R}_{knn}$ 称为论域 U 上的 k 近邻模糊关系.

基于 k 近邻模糊关系,研究了其作为模糊关系的基本性质.

性质 1. $\overline{R}_{knn}$ 是论域 U 上的自反关系.

性质 2. $\overline{R}_{knn}$ 在论域 U 上不满足对称性.

反例:不满足对称性的模糊关系 $\overline{R}_{1nn}$ 对应矩阵如下:

$$\boldsymbol{M}_{\overline{R}_{1nn}}=\begin{bmatrix}1 & 0 & 0.95 & 0\\ 0 & 1 & 0 & 1\\ 0.95 & 0.95 & 1 & 0\\ 0 & 1 & 0 & 0\end{bmatrix}\tag{5}$$

性质 3. $\overline{R}_{knn}$ 是论域 U 上不满足传递性.

分析:假设 $\overline{R}_{knn}$ 是论域 U 上的传递关系,则一定满足 $\overline{R}_{knn}\supseteq\overline{R}^2_{knn}$,即

$$R_{knn}(x_i,x_j)\geqslant\bigvee_{x_k\in U}[R_{knn}(x_i,x_k)\wedge R_{knn}(x_k,x_j)]\tag{6}$$

然而,给定如下 $|U|=4$ 的 $\overline{R}_{2nn}$,其对应矩阵如下:

$$\boldsymbol{M}_{\overline{R}_{2nn}}=\begin{bmatrix}1 & 0.9 & 0.95 & 0\\ 0 & 1 & 0.95 & 1\\ 0.95 & 0.95 & 1 & 0.8\\ 0.85 & 1 & 0 & 0\end{bmatrix}\tag{7}$$

其中 $\overline{R}_{2nn}(x_1,x_2)=0.9$.

$$\bigvee_{x_k\in U}[\overline{R}_{2nn}(x_1,x_k)\wedge\overline{R}_{2nn}(x_k,x_2)]$$

$$=(1\wedge 0.9)\vee(0.9\wedge 0.95)\vee(0.9.5\wedge 0.95)\vee(0\wedge 1)$$
$$=0.9\vee 0.9\vee 0.95\vee 0$$
$$=0.95,$$

则有 $0.9<0.95$,与公式(6)冲突.因此 $\overline{R}_{knn}$ 不满足传递性.

在此基础上,研究了 k 近邻模糊关系随 k 值变化呈现的规律.

定理 1. 给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,设 $|U|=n$.则对于任意的自然数 $k_1<k_2<n$,都有 $\overline{R}_{k_1nn}\subset\overline{R}_{k_2nn}$.

证明.根据定义 9,对于 $\forall x_i,x_j\in U$,$\overline{R}_{k_1nn}$ 和 $\overline{R}_{k_2nn}$ 可表示为

$$\overline{R}_{k_1m}=\begin{cases}\widetilde{R}(x_i,x_j), & x_j\in k_1nn(x_i)\\ 0, & x_j\in k_2nn(x_i)\wedge x_j\notin k_1nn(x_i)\\ 1, & i=j\\ 0, & \text{otherwise}\end{cases},$$

$$\overline{R}_{k_2m}=\begin{cases}\widetilde{R}(x_i,x_j), & x_j\in k_1nn(x_i)\\ \widetilde{R}(x_i,x_j), & x_j\in k_2nn(x_i)\wedge x_j\notin k_1nn(x_i)\\ 1, & i=j\\ 0, & \text{otherwise}\end{cases},$$

则 $\overline{R}_{k_1nn}$ 和 $\overline{R}_{k_2nn}$ 之间的关系分析可分为以下四种情况:

(1) 当 $x_j\in k_1nn(x_i)$ 时:

$\overline{R}_{k_1nn}(x_i,x_j)=\overline{R}_{k_2nn}(x_i,x_j)=\widetilde{R}(x_i,x_j)$；

(2) 当 $x_j \in k_2nn(x_i) \wedge x_j \notin k_1nn(x_i)$ 时：$\overline{R}_{k_1nn}(x_i,x_j)=0<\overline{R}_{k_2nn}(x_i,x_j)=\widetilde{R}(x_i,x_j)$；

(3) 当 $i=j$ 时：

$\overline{R}_{k_1nn}(x_i,x_j)=\overline{R}_{k_2nn}(x_i,x_j)=1$；

(4) 当其他情况时：

$\overline{R}_{k_1nn}(x_i,x_j)=\overline{R}_{k_2nn}(x_i,x_j)=0$，

因此，$\overline{R}_{k_1nn}(x_i,x_j)\leqslant\overline{R}_{k_2nn}(x_i,x_j)$ 恒成立.

且由情况(2)可知：当 $x_j \in k_2nn(x_i) \wedge x_j \notin k_1nn(x_i)$ 时，$\overline{R}_{k_1nn}(x_i,x_j) \neq \overline{R}_{k_2nn}(x_i,x_j)$. 因此，$\overline{R}_{k_1nn} \neq \overline{R}_{k_2nn}$.

综上，$\overline{R}_{k_1nn} \subset \overline{R}_{k_2nn}$ 得证.

定理 2. 给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$，设 $|U|=n$. 则当 $k=n-1$ 时，$\overline{R}_{knn}$ 为论域 U 上的模糊相似关系，即 $\overline{R}_{knn}=\widetilde{R}$ 成立.

证明. 当 $k=n-1$ 时，为避免歧义，论域 U 上的 $n-1$ 近邻模糊关系简记为 $\overline{R}_{n-1}$. 由定义 9 可知，对于 $\forall x_i,x_j \in U$，存在以下两种情况：

(1) 当 $i=j$ 时：

$\overline{R}_{n-1}(x_i,x_j)=\widetilde{R}(x_i,x_j)=1$；

(2) 当 x_j 是 x_i 的 $n-1$ 近邻时：

$\overline{R}_{n-1}(x_i,x_j)=\widetilde{R}(x_i,x_j)$.

同时，由于 x_i 的 $n-1$ 近邻至少包含 $n-1$ 个对象，因此对于 $\forall x_i,x_j \in U$，情况(1)和(2)已包含 x_i 与论域中其余所有 $n-1$ 个对象. 则不存在定义 9 中的“otherwise”的情况.

综上所述，当 $k=n-1$ 时，$\overline{R}_{knn}=\widetilde{R}$ 得证.

定理 1 和定理 2 表明：随着 k 的逐渐增大，k 近邻模糊关系将形成一个模糊关系套，当 $k=n-1$ 时，k 近邻模糊关系退化为模糊相似关系.

3.1.2 k 近邻模糊关系及其性质

从图 3 可知，构建论域中对象间模糊关系的目的是提取有效值和有效位置，因此对 k 近邻模糊关系进行参数分析的最终目的是找出使得 k 近邻模糊关系能够包含所有有效值的最小 k 值.

经过大量实例和理论分析，我们得到以下结论.

定理 3. 给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$，令 $\overline{R}_{2nn}$ 为论域 U 上的 2 近邻模糊关系，$\widetilde{R}$ 为 U 上的模糊相似关系，则对于任意的 $y \in Eff(\widetilde{R})$，$y \in \overline{R}_{2nn}$.

证明. 根据公式 9，对于 $\forall x_i,x_j \in U$，$\overline{R}_{2nn}$ 可表示为

$$\overline{R}_{2nn}=\begin{cases}\widetilde{R}(x_i,x_j), & x_j \in 2nn(x_i)\\ 1, & i=j\\ 0, & \text{otherwise}\end{cases} \tag{8}$$

根据图 3，基于模糊相似关系的有效值提取分为两个阶段：

(1) 除主对角线元素外，模糊相似关系的行最大值一定为有效值. 对于 $\forall x_i \in U$，设 $y=\widetilde{R}(x_i,x_k)$ 为第 x_i 行除主对角线元素外的最大值，即

$$y=\max_{1\leqslant j\leqslant n,j\neq i}(\widetilde{R}(x_i,x_j)),$$

则 y 是 x_i 的最近邻，所以 $y \in 2nn(x_i)$ 一定成立，根据公式(3.5)，$\overline{R}_{2nn}(x_i,x_k)=\widetilde{R}(x_i,x_k)=y$. 因此，$y \in \overline{R}_{2nn}$.

(2) 除主对角线元素和行最大值外，依次取模糊相似关系中的最大值作为有效值. 设选择最大值后，会依次将 $\widetilde{R}$ 中最大值所在位置为 0. 设第 $z-1$ 轮选择最大值后，得到的模糊关系为 $\widetilde{R}_{z-1}$. 令第 z 轮最大值为 $\max_z$，则存在以下两种情况：

(a) $\max_z$ 为 x_i 的第 2 近邻，则 $\max_z \in \overline{R}_{2nn}$ 成立；

(b) $\max_z$ 不为论域中任何一个对象的第 2 近邻对应值，假设 $\max_z=\widetilde{R}(x_i,x_j)$. 由于模糊相似关系具有对称性，则 $\max_z=\widetilde{R}(x_i,x_j)=\widetilde{R}(x_j,x_i)$.

因为 $\max_z$ 是第 z 轮的最大值，所以 $\max_z=\max\widetilde{R}_{z-1} \geqslant \widetilde{R}(x_i,x_j^2)$. 由于该阶段为第二阶段，则 $\max_z=\widetilde{R}(x_i,x_j^2)$. 则 $\max_z \in \overline{R}_{2nn}$ 成立.

综上所述，对于任意的 $y \in Eff(\widetilde{R})$，$y \in \overline{R}_{2nn}$ 得证.

根据定理 3，可以得出：通过模糊商空间模型构建多粒度空间时，仅需要构建 2 近邻模糊关系，即可包含模糊相似关系中的全部有效信息，与对应的模糊等价关系具有相同的分类能力，构建出完全相同的多粒度空间.

为方便后续研究，基于 2 近邻模糊关系，定义了论域的最近邻序列和次近邻序列.

定义 10. 最近邻序列和次近邻序列. 给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$. 则论域 U 的最近邻序列 $1st(U)$ 和次近邻序列 $2nd(U)$ 定义为

$$\begin{aligned}1st(U)&=\{\langle(x,x^1),\widetilde{R}(x,x^1)\rangle \mid x \in U\}\\ 2nd(U)&=\{\langle(x,x^2),\widetilde{R}(x,x^2)\rangle \mid x \in U\}\end{aligned} \tag{9}$$

其中,x^1 代表对象 x 的第 1 近邻(不含自身),x^2 代表对象 x 的第 2 近邻(不含自身).为描述方便,1st(U) 和 2nd(U) 简写为 1st 和 2nd.

3.2　有效值和有效位置的提取

由图 3 可知,基于模糊相似关系的有效值和有效位置的提取,只有满足以下两个条件才能够搜索出全部的有效值和有效位置:

(1) 模糊相似关系对应矩阵的上三角区域中所有行或列至少存在一个有效位置;

(2) 模糊相似关系对应矩阵的上三角区域中有效位置的数目 $\geqslant n-1$.

但是由于 2 近邻模糊关系不具有对称性,因此需要重新设计有效值和有效位置的提取算法.本节将基于 2 近邻模糊关系,针对最近邻序列和次近邻序列,设计两阶段的有效值和有效位置的提取算法.

首先提出最近邻阶段和次近邻阶段的有效位置数计算方法.

特别地,2 近邻模糊关系虽然不具有对称性,但论域中可能存在部分对象互为 2 近邻,因此关系内部仍然可能存在值的对称.也就是说,$\overline{R}_{2nn}(x_i,x_j)=\overline{R}_{2nn}(x_j,x_i)\neq 0$.如公式(3.5)中的 $\overline{R}_{2nn}(x_1,x_3)=\overline{R}_{2nn}(x_3,x_1)=0.95$,$x_1$ 和 x_3 互为最近邻,在模糊关系中占据两个位置,但是它们对划分论域、构造多粒度空间的贡献完全一样,属于冗余信息,计算有效位置时需要去除.因此给出以下定义:

定义 11.相互 2 近邻.给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,令 $\overline{R}_{2nn}$ 为 2 近邻模糊关系.对于 $\forall x_i,x_j\in U(x_i\neq x_j)$,若满足

$$\overline{R}_{2nn}(x_i,x_j)=\overline{R}_{2nn}(x_j,x_i)\neq 0,$$

则 x_i 和 x_j 为相互 2 近邻,记作 $m2nn(x_i,x_j)$.

(1) 最近邻阶段的有效位置

模糊相似关系有效值的提取需要首先提出除主对角线元素外的行(列)最大值,再对应搜索其有效位置.因此基于 2 近邻模糊关系,第一阶段仅需搜索最近邻序列对应的有效位置即可.

基本思想:最近邻阶段的有效位置数=最近邻序列的势+次近邻序列中取到最近邻有效值的位置数−最近邻有效值的相互 2 近邻数.具体形式化表达如下:

基于定义 11,最近邻阶段的相互 2 近邻数为

$$\begin{aligned}&m2nn(1\mathrm{st})\\&=|\{\overline{R}_{1nn}(x_i,x_j)\,|\,m2nn(x_i,x_j),i<j\}|\end{aligned}$$

最近邻序列对应的有效位置不仅包含最近邻序列的所有位置,同时包含次近邻序列中取到最近邻阶段有效值的所有位置.因此给出以下定义:

定义 12.给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,论域 U 的最近邻序列为 $1st$,次近邻序列为 $2nd$.对于 $\forall<(x,x^2),\widetilde{R}(x,x^2)>\in 2nd$,若 $\exists\langle(y,y^1),\widetilde{R}(y,y^1)\rangle\in 1st$,使得

$$\widetilde{R}(y,y^1)=\widetilde{R}(x,x^2),$$

则 $\langle(x,x^2),\widetilde{R}(x,x^2)\rangle$ 为次近邻序列中取到最近邻阶段有效值的有效位置.包含所有该有效位置的集合记为 $Eff_{p1}(2nd)$.

基于上述分析,最近邻阶段的有效位置数计算如下:

定义 13.最近邻阶段的有效位置数.给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,论域 U 的最近邻序列为 $1st$ 且 $\overline{R}_{2nn}$ 为 2 近邻模糊关系.则最近邻阶段的有效位置数为

$$\begin{aligned}&|Eff_{p1}(\overline{R}_{2nn})|\\&=|1st|+|Eff_{p1}(2\mathrm{nd})|-|m2nn(1\mathrm{st})|\end{aligned}\tag{10}$$

其中 $|1\mathrm{st}|$ 代表最近邻序列的势,$|Eff_{p1}(2\mathrm{nd})|$ 代表次近邻序列中取到最近邻阶段有效值的有效位置数目,$|m2nn(1\mathrm{st})|$ 代表最近邻阶段有效值的相互 2 近邻数.

注:相互近邻数可由 k 近邻算法直接输出.

(2) 次近邻阶段的有效位置数

模糊相似关系有效值的提取需要在提取除主对角线元素外行(列)最大值之后,仍然需要依次提取剩余值的最大值,直到有效位置满足条件.因此基于 2 近邻模糊关系,第二阶段需要依次提取次近邻序列的最大值及其对应的有效位置.

基本思想为:次近邻阶段的有效位置数=现有的有效位置数+当前最大值在次近邻序列中的频数.形式化表达如下:

定义 14.次近邻阶段的有效位置数.给定一个模糊决策信息系统 $FDIS=<U,At,V,f>$,论域 U 的次近邻序列为 2nd 且 $\overline{R}_{2nn}$ 为 2 近邻模糊关系.设当前为次近邻阶段第 i 次取值,当前次近邻序列中的最大值为 $\max_i$,则次近邻阶段第 i 次的有效位置数为

$$\begin{aligned}&|Eff_{p2}(\overline{R}_{2nn},i)|\\&=\begin{cases}|Eff_{p1}(\overline{R}_{2nn})|+\mu(\max_i), & i=1\\|Eff_{p2}(\overline{R}_{2nn},i-1)|+\mu(\max_i), & i>1\end{cases}\end{aligned}\tag{11}$$

其中，$\mu(\max_i)$ 代表 $\max_i$ 的频数.

据上述分析，基于 2 近邻模糊关系提取有效值和有效位置算法，如图 4 所示.

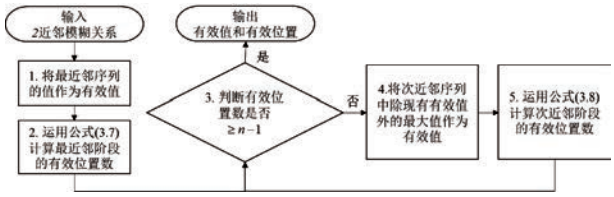


图 4 2 近邻模糊关系的有效值和有效位置提取

3.3 多粒度空间快速构建方法

本节将基于 2 近邻模糊关系提出多粒度空间快速构建方法，3.3.1 节将进行算法的设计，3.3.2 节将分析该算法的时间复杂度，并同现有算法进行对比；3.3.3 节将针对基于 2 近邻模糊关系提出多粒度空间快速构建方法，进行实例分析，展示该算法的实施流程.

3.3.1 算法设计

根据 3.1 和 3.2 节的理论分析，基于 2 近邻模糊关系的多粒度空间快速构建方法分为 2 近邻模糊关系、有效值和有效位置、截关系和多粒度空间四个步骤，如图 5 所示. 其中，步骤 2 根据图 4 所示算法流程，从 2 近邻模糊关系中提取有效值和有效位置.

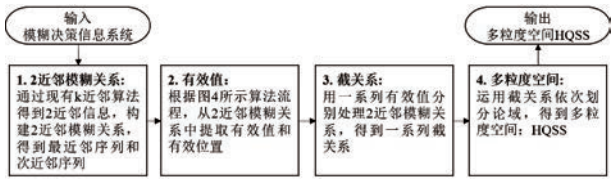


图 5 基于 2 近邻模糊关系的多粒度空间快速构建

具体的，如算法 1 所示. 相比现有的经典模糊商空间模型和基于有效值的模糊商空间模型构建多粒度空间，本文方法的优势在于：

(1) 构建了只包含模糊相似关系中稀疏有效信息的 2 近邻模糊关系，避免了大量无效相似性度量；

(2) 基于 2 近邻模糊关系设计了对应的有效值和有效位置的快速提取算法，避免了 2 近邻模糊关系求传递闭包的时间损耗；

(3) 运用 p 个有效值分别和最多包含 $2n$ 个值的 2 近邻模糊关系进行对比，直接构建截关系，相比现有方法需要 k 个有效值分别和包含 n^2 个值的模糊相似关系或模糊等价关系进行对比，提升了该步骤的效率.

算法 1. 基于 2 近邻模糊关系的多粒度空间快速构建.

输入：模糊决策信息系统 $FDIS=<U,At,V,f>$；

输出：多粒度空间 $HQSS$.

//初始化：$|Eff(\overline{R}_{2nn})|\leftarrow\phi$，$|Eff_p(\overline{R}_{2nn})|\leftarrow\phi$，$i\leftarrow 1$；

//步骤 1：2 近邻模糊关系；

运用 k 近邻算法得到：$\overline{R}_{2nn}$，$1sd$，$2nd$；

//步骤 2：有效值和有效位置；

//步骤 2.1：提取最近邻阶段的有效值，计算有效位置数；

有效值集合：$Eff(\overline{R}_{2nn})=\{\overline{R}_{2nn}(x_i,x_j)|\langle(x_i,x_j),\overline{R}_{2nn}(x_i,x_j)\rangle\in 1st\}$；

根据公式(3.7)计算 $|Eff_{p1}(\overline{R}_{2nn})|$；

有效位置数：$|Eff_p(\overline{R}_{2nn})|=|Eff_{p1}(\overline{R}_{2nn})|$；

WHILE $|Eff_p(\overline{R}_{2nn})|<n-1$；

//步骤 2.2：提取次近邻阶段的有效值，计算有效位置数；

$A=\{\overline{R}_{2nn}(x_i,x_j)|\langle(x_i,x_j),\overline{R}_{2nn}(x_i,x_j)\rangle\in 2nd\}$；

$i\leftarrow i+1$；

提取当前 $2nd$ 的最大值：$max=\max A$；

有效值集合：$Eff(\overline{R}_{2nn})=Eff(\overline{R}_{2nn})\cup\{max\}$；

根据公式(3.8)计算 $|Eff_{p2}(\overline{R}_{2nn})|$；

有效位置数：$|Eff_p(\overline{R}_{2nn})|=|Eff_{p2}(\overline{R}_{2nn})|$；

END WHILE

//步骤 3：截关系；

FOR λ IN $Eff(\overline{R}_{2nn})$；

FOR $\overline{R}_{2nn}(x,y)$ **IN** $1sd\cup 2nd$；

IF $\lambda>\overline{R}_{2nn}(x,y)$；

$\overline{R}_{2nn,\lambda}(x,y)=1$；

ELSE

$\overline{R}_{2nn,\lambda}(x,y)=0$；

END IF

END FOR

END FOR

//步骤 4：多粒度空间；

FOR λ IN $Eff(\overline{R}_{2nn})$；

运用 $\overline{R}_{2nn,\lambda}$ 划分论域，得到 $\pi_{\overline{R}_{2nn,\lambda}}(U)$；

$\pi_{\overline{R}_{2nn}}(U)\leftarrow\pi_{\overline{R}_{2nn}}(U)\cup\{\pi_{\overline{R}_{2nn,\lambda}}(U)\}$；

END FOR

RETURN $\pi_{\overline{R}_{2nn}}(U)$.

3.3.2 时间复杂度分析

算法 1 的时间复杂度由四个步骤组成. 步骤 1 是构建 2 近邻模糊关系，步骤 2 是提取有效值和有效位置，步骤 3 是构建截关系，步骤 4 是构建多粒度空间. 所有步骤的时间复杂度分析如下：

(1)对于步骤 1,本文使用 KD-Tree 构建 2 近邻模糊关系,时间复杂度为 $O(n\log n)$.现有方法均是通过计算论域中两两对象之间的相似度构建模糊相似关系,时间复杂度为 $O(n^2)$,因此,本文方法降低了步骤 1 的时间复杂度;

(2)对于步骤 2,本文按照图 4 的步骤设计算法提取 2 近邻模糊关系中的有效值和有效位置,分如下 3 种情况讨论时间复杂度:

①最好情况:若只进行最近邻阶段即可提取出全部的有效值,则时间复杂度为 $O(1)$;

②最差情况:若全部遍历完次近邻序列的所有值,则需要对次近邻序列进行排序,时间复杂度为 $O(n\log n)$;

③一般情况:设次近邻阶段的循环次数为 m,则步骤 2 的一般时间复杂度为 $O(mn)$.

经典模糊商空间模型通过平方法计算传递闭包提取有效值的时间复杂度为 $O(n^5)$,经典模糊商空间模型通过 Warshall 算法计算传递闭包提取有效值的时间复杂度为 $O(n^3)$,基于有效值的模糊商空间模型提取有效值的时间复杂度为 $O(n^2)$,因此,本文方法降低了步骤 2 的时间复杂度;

(3)对于步骤 3,本文使用所有有效值分别构建 2 近邻模糊关系的截关系,设存在 p 个有效值需要对比 2 近邻模糊关系中的 $2n$ 个值,构建 p 个截关系,则时间复杂度为 $O(2pn)$.经典模糊商空间模型通过 p 个有效值对比模糊等价关系中 $2n$ 个值,基于有效值的模糊商空间模型通过 p 个有效值对比模糊相似关系中 $2n$ 个值,构建 p 个截关系,时间复杂度均为时间复杂度为 $O(pn^2)$.因此,本文方法降低了步骤 3 的时间复杂度;

(4)对于步骤 4,本文 p 个截关系,最多需要对比截关系中的 $2n$ 个值来划分论域,形成 p 层的多粒度空间,时间复杂度最差为 $O(2pn)$.现有方法均是通过对比截关系中的 n^2 个值来划分论域形成 p 层的多粒度空间,时间复杂度为 $O(pn^2)$,因此,本文方法降低了步骤 4 的时间复杂度.

本文方法和现有方法的时间复杂度对比如图 6 所示,图 6 中将时间复杂度划分为 6 个等级,如 $O(n^5)$ 对应等级 6,$O(n^3)$ 对应等级 5 等.等级越低,该步骤的时间复杂度越低.由图 6 可知,与现有方法构建多粒度空间的时间复杂度相比,本文方法在所有步骤的效率都有较大的提升.

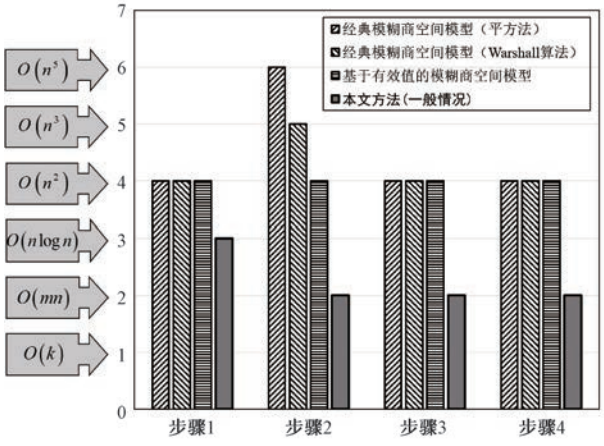


图 6　时间复杂度对比

3.3.3　实例分析

为了更加清晰地阐述和展示基于 2 近邻模糊关系的模糊商空间模型来快速构建多粒度空间的流程,本节将给出详细的实例分析.

实例 1 给定一个种子发芽试验的模糊决策信息系统 $FDIS=<U,At,V,f>$,令 $U=\{x_1,x_2,\cdots,x_7\}$ 代表包含 1 个种子样本的论域,$At=C\cup D$,$C=\{r_1,r_2,\cdots,r_6\}$ 代表条件属性集,$D=\{d\}$ 代表决策属性集.属性 r_1 代表天气,是集合{晴天,多云,雨天}上的模糊集;属性 r_2 代表湿度,是集合{潮湿,中性,干燥}上的模糊集;属性 r_3 代表温度,是集合{高,平均,低}上的模糊集;属性 r_4 代表照明,是集合{强,适中,弱}上的模糊集;属性 r_5 代表土壤条件,是集合{肥沃,一般,贫瘠}上的模糊集;属性 r_6 代表是否使用切片,0 代表未使用,0 代表使用;决策属性 d 种子发芽与否,0 代表未发芽,1 代表发芽.该模糊决策信息系统的具体信息展示在表 3 中.

(1)2 近邻模糊关系:基于论域中所有对象的 2 近邻信息,根据定义 9 构建 2 近邻模糊关系,其对应矩阵如下:

$$\boldsymbol{M}_{R_{2m}}=\begin{bmatrix}1 & 0 & 0.53 & 0 & 0.57 & 0 & 0\\ 0 & 1 & 0.53 & 0 & 0 & 0 & 0.47\\ 0.53 & 0.53 & 1 & 0 & 0 & 0.57 & 0\\ 0.47 & 0 & 0 & 1 & 0.53 & 0 & 0\\ 0.57 & 0 & 0 & 0.53 & 1 & 0 & 0\\ 0.50 & 0 & 0.57 & 0 & 0.50 & 1 & 0\\ 0.47 & 0.47 & 0.47 & 0 & 0.43 & 0.43 & 1\end{bmatrix}.$$

并根据定义 10 得到相应的最近邻序列和次近邻序列:

表 3 模糊决策信息系统

	r_1	r_2	r_3	r_4	r_5	r_6	d
x_1	(0.8,0.2,0.0)	(0.2,0.8,0.1)	(0.1,0.7,0.4)	(0.1,0.6,0.2)	(0.1,0.8,0.1)	1	1
x_2	(0.9,0.1,0.0)	(0.1,0.9,0.1)	(0.0,0.2,0.8)	(0.1,0.0,0.0)	(0.4,0.6,0.1)	1	0
x_3	(0.7,0.3,0.0)	(0.8,0.2,0.1)	(0.2,0.5,0.2)	(0.9,0.2,0.0)	(0.3,0.7,0.2)	0	1
x_4	(0.0,0.3,0.7)	(0.9,0.2,0.1)	(0.0,0.3,0.8)	(0.0,0.2,0.9)	(0.5,0.3,0.1)	0	0
x_5	(0.2,0.8,0.1)	(0.2,0.9,0.1)	(0.1,0.9,0.1)	(0.0,0.7,0.1)	(0.2,0.6,0.0)	1	1
x_6	(0.8,0.3,0.1)	(0.1,0.8,0.1)	(0.8,0.3,0.0)	(0.8,0.1,0.0)	(0.7,0.2,0.1)	0	1
x_7	(0.2,0.7,0.1)	(0.0,0.2,0.9)	(0.0,0.2,0.8)	(0.9,0.2,0.0)	(0.4,0.5,0.1)	1	0

$$
\begin{aligned}
&1st(U)\\
=&\{\langle(x_1,x_5),0.57\rangle,\langle(x_2,x_3),0.53\rangle,\langle(x_3,x_6),0.57\rangle,\\
&\langle(x_4,x_5),0.53\rangle,\langle(x_5,x_1),0.57\rangle,\langle(x_6,x_3),0.57\rangle,\\
&\langle(x_7,x_1),0.47\rangle,\langle(x_7,x_2),0.47\rangle,\langle(x_7,x_3),0.47\rangle\}\\
&2nd(U)\\
=&\{\langle(x_1,x_3),0.53\rangle,\langle(x_2,x_7),0.47\rangle,\langle(x_3,x_1),0.53\rangle,\\
&\langle(x_3,x_2),0.53\rangle,\langle(x_4,x_1),0.47\rangle,\langle(x_5,x_4),0.53\rangle,\\
&\langle(x_6,x_1),0.50\rangle,\langle(x_6,x_5),0.40\rangle,\langle(x_7,x_5),0.43\rangle,\\
&\langle(x_7,x_6),0.43\rangle\}.
\end{aligned}
$$

(2) 有效值和有效位置:将最近邻序列中的值提取为有效值,则有 $\{0.47,0.53,0.57\}$. 当前处于最近邻阶段,该阶段存在的相互 2 近邻为

$$
\begin{aligned}
&|m2nn(1st)|\\
=&|\{\overline{R}_{1nn}(x_i,x_j)\mid m2nn(x_i,x_j),i<j\}|\\
=&|\{\overline{R}_{1nn}(x_1,x_5),\overline{R}_{1nn}(x_2,x_3),\overline{R}_{1nn}(x_3,x_6)\\
&\overline{R}_{1nn}(x_4,x_5),\overline{R}_{1nn}(x_2,x_7)\}|\\
=&5.
\end{aligned}
$$

根据公式(10)次近邻序列中取到最近邻阶段有效值的有效位置数目为

$$
\begin{aligned}
&|Eff_{p1}(2nd)|\\
=&|\{\langle(x_1,x_3),0.53\rangle,\langle(x_2,x_7),0.47\rangle,\langle(x_3,x_1),0.53\rangle,\\
&\langle(x_3,x_2),0.53\rangle,\langle(x_4,x_1),0.47\rangle,\langle(x_5,x_4),0.53\rangle\}|\\
=&6.
\end{aligned}
$$

因此,根据公式(11),最近邻阶段的有效位置数为

$$
\begin{aligned}
&|Eff_{p1}(\overline{R}_{2nn})|\\
=&|1st|+|Eff_{p1}(2nd)|-|m2nn(1st)|\\
=&9+6-5\\
=&10>6.
\end{aligned}
$$

因此,$|Eff(U)|=\{0.47,0.53,0.57\}$.

(2) 截关系:根据算法 1 的步骤 3,运用 $Eff(U)$ 中的值分别处理 2 近邻模糊关系,可以得到一系列截关系:四个截关系及其相应的截矩阵 $\boldsymbol{M}_{\overline{R}_{2nn,\lambda_1}}$,$M_{\overline{R}_{2nn,\lambda_2}}$,$M_{\overline{R}_{2nn,\lambda_3}}$,$M_{\overline{R}_{2nn,\lambda_4}}$,由于篇幅原因,只给出 $M_{\overline{R}_{2nn,\lambda_3}}$:

$$
\boldsymbol{M}_{\overline{R}_{2nn,\lambda_3}}=\begin{bmatrix}1&0&0&0&1&0&0\\0&1&0&0&0&0&0\\0&0&1&0&0&1&0\\0&0&0&1&0&0&0\\1&0&0&0&1&0&0\\0&0&0&0&1&1&0\\0&0&0&0&0&0&1\end{bmatrix}.
$$

(4) 多粒度空间:根据算法 1 的步骤 4,根据 4 个截矩阵逐一划分论域,形成多粒度空间:

$$
\begin{aligned}
&\pi_{\overline{R}_{2nn,\lambda_1}}(U)=\{\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}\},\\
&\pi_{\overline{R}_{2nn,\lambda_2}}(U)=\{\{x_1,x_2,x_3,x_4,x_5,x_6\},\{x_7\}\},\\
&\pi_{\overline{R}_{2nn,\lambda_3}}(U)=\{\{x_1,x_5\},\{x_2\},\{x_4\},\{x_3,x_6\},\{x_7\}\},\\
&\pi_{\overline{R}_{2nn,\lambda_3}}(U)=\{\{x_1\},\{x_2\},\{x_3\},\{x_4\},\{x_5\},\{x_6\},\{x_7\}\}.
\end{aligned}
$$

4 实验分析

相比于现有的模糊商空间模型构建多粒度空间,基于 2 近邻模糊关系的多粒度空间快速构建方法通过构建只包含有效信息的 2 近邻模糊关系,极大地降低了各个步骤的时间复杂度. 为了验证所提方法的贡献,本节将通过实验进行验证.

实验配置

(1) 随机选择 7 个来源于公开的 UCI 数据集学习库(https://archive.ics.uci.edu/ml/index.php)中的数据集,其详细信息如表 4 所示.

表 4 UCI 数据集的详细信息

编号	UCI 数据集	对象数	属性值	类别数
1	Iris	150	3	3
2	Heart	270	12	2
3	Ecoli	336	7	8
4	Breast-cancer-wisconsin	699	8	2
5	Banknote	1372	4	2
6	Titanic	2200	5	2
7	PhishingData	2456	29	3
8	Gender Gap in Spanish WP	4746	20	3
9	Wavaform	5000	20	3

(2)3 个来源于英国的生物库医学数据库(UKB)中的数据集,其详细信息如表 5 所示;

表 5　UKB 数据集的详细信息

编号	UKB 数据集	对象数	属性值	类别数
10	UKB1	2000	459	2
11	UKB2	2000	766	2
12	UKB3	2000	508	2

数据可用性声明:本实验中使用的 UKB 数据集来自用户 id:51470. 研究人员可以通过官方网站(https://www.ukbiobank.ac.uk)申请使用 UKB 数据.

(3) 随机选择 3 个图像数据集,其详细信息如表 6 所示;

表 6　图像数据集的详细信息

编号	图像数据集	对象数	属性值	类别数
13	ORL	400	10245	40
14	COIL20	1440	1024	20
15	COIL100	2000	508	100

(4) 3 个来源于网址(https://jundongl.github.io/scikit-feature/datasets.html)的文本数据集,其详细信息如表 7 所示.

表 7　文本数据集的详细信息

编号	文本数据集	对象数	属性值	类别数
16	RELATHE	1427	4322	2
17	PCMA	1943	3829	2
18	BASEHOCK	1993	4862	2

实验结果与分析

本节将通过以下四个方面的实验来验证所提方法:

(1)效率. 4.1 节将通过对比经典模糊商空间模型、基于有效值的模糊商空间模型和本文方法的多粒度空间构建时间,验证本文方法构建多粒度空间的效率;

(2)正确性. 4.2 节将通过对比经典模糊商空间模型、基于有效值的模糊商空间模型和本文方法构建出的多粒度空间,验证本文方法构建多粒度空间的正确性;

(3)有效性. 4.3 节将本文所提多粒度空间快速构建方法应用于分类任务,与粒计算分类器和经典机器学习分类器对比,验证本文方法面向分类任务的有效性、鲁棒性和稳定性;

(4)显著性. 4.4 节将本文方法应用于分类任务的表现与粒计算分类器、经典机器学习分类器之间的统计分析,验证本文方法面向分类任务的显著性.

4.1　构建多粒度空间的效率评估

在本节中,将通过对比本文方法(基于 2 近邻模糊关系的模糊商空间模型)与现有方法(经典模糊商空间模型(平方法)、经典模糊商空间模型(Warshall 算法)、基于有效值的模糊商空间模型)构建多粒度空间的时间来验证本文方法在效率方面的有效性. 实验结果如表 8 所示.

表 8　多粒度空间构建时间对比

编号	本文方法	基于有效值的模糊商空间模型	经典模糊商空间模型(平方法)	经典模糊商空间理模型(Warshall 算法)
1	0.0938	0.2031	3.2146	6.2061
2	0.2818	0.9078	13.4869	31.3979
3	0.5955	1.5498	32.4581	69.8845
4	0.5475	14.7773	282.2334	601.0443
5	8.4793	31.4319	899.1458	3552.8956
6	10.5686	50.5568	1954.3153	6023.9564
7	0.0313	2.7849	2460.5974	6591.4598
8	10.9569	116.3888	5846.1239	15306.5914
9	20.7037	132.7361	10996.1235	17269.8546
10	0.1412	25.6333	2023.4569	5923.6855
11	0.1875	29.9750	2245.9537	6915.2655
12	0.1563	26.2730	1965.4863	6002.9990
13	0.0963	0.1346	50.0236	78.6692
14	0.1875	0.9991	45.3386	150.1990
15	0.2322	30.4578	3464.5841	9889.9990
16	0.3632	35.5469	2364.5695	8889.4585
17	0.4895	30.8945	2956.5841	8150.1990
18	0.4596	32.9638	2546.4595	9323.4585

如表 8 所示，在 9 个 UCI 数据集、3 个 UKB 数据集、3 个图像数据集和 3 个文本数据集中，本文方法构建多粒度空间的运行时间远低于现有三种模糊商空间模型构建多粒度空间的运行时间. 因此，验证了本文方法在时间复杂度方面的优势. 基于 2 近邻模糊关系，多粒度空间的构建效率得到了极大的提升.

由于现有方法构建多粒度空间均需要构建模糊相似关系，因此本文方法构建多粒度空间效率最高的原因分析如下：

(1) 本文使用 KD-Tree 算法构建了包含了模糊相似关系中全部有效值的 2 近邻模糊关系，因此在构建多粒度空间的步骤 1 中，避免了构建模糊相似关系时计算论域中两两之间相似度的计算复杂度，在一定程度上提升了本文方法的效率；

(2) 本文构建的 2 近邻模糊关系对应模糊矩阵的行只包含 2 近邻信息，因此在构建多粒度空间的步骤 2 中，避免了从包含大量冗余信息的模糊相似关系中提取有效值的大量计算复杂度，在一定程度上提升了本文方法的效率；

(3) 本文构建的 2 近邻模糊关系只包含 $2n$ 个值，因此在构建多粒度空间的步骤 3 和 4 中，若有效值个数为 p 时只需要进行 $2pn$ 次对比，避免了现有方法需要对比 $n \times n$ 矩阵中所有值的计算复杂度，在一定程度上提升了本文方法的效率.

4.2 构建多粒度空间的正确性评估

在本节中，将通过对比本文方法(基于 2 近邻模糊关系的模糊商空间模型)与现有方法(经典模糊商空间模型(平方法)、经典模糊商空间模型(Warshall 算法)、基于有效值的模糊商空间模型)所提取的有效信息和构建多粒度空间的相似度，来验证本文方法提取有效信息和构建多粒度空间的正确性. 实验结果如表 9 所示.

表 9 提取有效值和构建多粒度空间的正确性

编号	有效信息个数				本文方法提取有效值的正确性	本文方法构建多粒度空间的正确性
	本文方法	基于有效值的模糊商空间模型	经典模糊商空间模型(平方法)	经典模糊商空间理模型(Warshall 算法)		
1	21	21	21	21	100%	100%
2	249	249	249	249	100%	100%
3	311	311	311	311	100%	100%
4	45	45	45	45	100%	100%
5	956	956	956	956	100%	100%
6	3	3	3	3	100%	100%
7	5	5	5	5	100%	100%
8	1394	1394	1394	1394	100%	100%
9	981	981	981	981	100%	100%
10	1490	1490	1490	1490	100%	100%
11	1488	1488	1488	1488	100%	100%
12	1485	1485	1485	1485	100%	100%
13	366	366	366	366	100%	100%
14	1439	1439	1439	1439	100%	100%
15	7154	7154	7154	7154	100%	100%
16	95	95	95	95	100%	100%
17	323	323	323	323	100%	100%
18	1125	1125	1125	1125	100%	100%

如表 9 所示，在 9 个 UCI 数据集、3 个 UKB 数据集、3 个图像数据集和 3 个文本数据集中，本文方法与现有三种模糊商空间模型提取有效信息的数目均完全一致，有效信息完全相同. 同时，本文方法与现有三种模糊商空间模型构建的多粒度空间也完全相同. 因此，验证了本文方法在构建多粒度空间方面的正确性. 本文提出基于 2 近邻模糊关系构建多粒度空间的方法，不仅提升了构建效率，同时保证了构建质量.

4.3 面向分类任务的有效性评估

在本节中，本文所提多粒度空间快速构建方法将被应用到用分类任务中，并验证其有效性.

基于本文方法构建多粒度空间并用于分类任务的流程如图 7 所示，具体步骤及其时间复杂度如下：

(1) 数据预处理：使用平均值法进行数据集的缺失值填充，使用极差法进行数据集的标准化；

(2) 构建多粒度空间：根据算法 1 构建多粒度空间. 该步骤的时间复杂度如图 6 所示为 $O(n\log n)$；

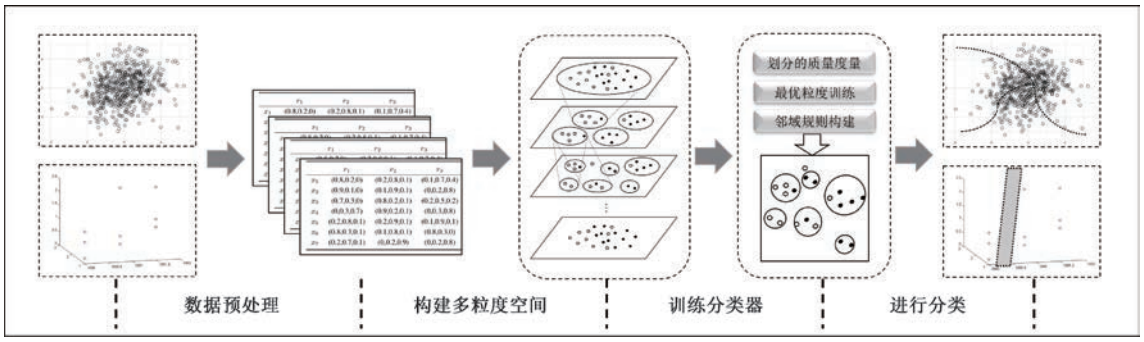


图 7　基于本文方法构建多粒度空间用于分类的流程

(3) 训练分类器:将数据集随机分为训练集和测试集,运用文献[42]中划分的质量度量方法,输入训练集学习多粒度空间中"质量最高"的最优知识粒度,并在最优知识粒度上构建邻域规则.该步骤的时间复杂度为 $O(p+1)$,其中 p 为有效信息的个数,$p+1$ 为构建出的多粒度空间的层数;

(4) 进行分类:输入测试集,运用文献[42]中的分类策略并运用中心点偏移量处理决策冲突,实现分类.该步骤的时间复杂度为 $O(q)$,其中 q 为步骤(3)中选择的最优粒度中的规则数.

综上所述,基于本文方法构建多粒度空间并用于分类任务的时间复杂度为 $O(n\log n)$.

为了评估本文方法的性能,实验采用 10 折交叉验证进行,采用评估分类性能的 *Accuracy*、*Recall*、*Precision* 和 $F1$ 四个通用指标测试其性能,采用标准差(STD)来度量分类器的稳定性.同时将比较以下两种类型的分类器:

(1) 9 个粒计算分类器:NCR[60],NCRDPP[61],TNCR[62],GBNRS[63],GB-kNN[64],FNC-TC[65],FNC-EC[66],RSLRS[67]和 E3WD[68];

(2) 6 个经典机器学习分类器:决策树(Classification and Regression Tree, CART)[69]、支持向量机(Support Vector Machine, SVM)[70]、朴素贝叶斯(Naive Bayes Classifier, NBC)[71]、逻辑回归分类器(Logistic Regression classifier, LR)[72]、随机森林(Random forest, RF)[73] 和 k 近邻(k-Nearest Neighbors, kNN)[74].

分类器的参数设置如表 10 所示.

对于 16 个分类器,在 18 个数据集上关于四个分类指标 *Accuracy*、*Recall*、*Precision* 和 $F1$ 的实验结果如表 11～14 所示,在四个指标下的平均性能如图 8.表 11～14 中第一行是分类器,第一列是数据集的编号,表格中的数据代表不同分类器在不同数据集上的分类性能,性能指标取值范围为 0 到 1,值越大分类器性能越好;表中的 STD 为 18 个数据集分类性能的标准差,STD 越小说明分类器性能越稳定.

表 10　分类器的参数设置

分类器	参数设置
NCR	使用文献[60]中的相同参数
NCRDPP	使用文献[61]中的相同参数
TNCR	使用文献[62]中的相同参数
GBNRS	粒球纯度的阈值为 1.
CB-kNN	(1) 粒球纯度的阈值为 1; (2) 个努力求自适应 k 的取值.
FNC-TC	使用文献[65]中的相同参数
FNC-EC	使用文献[66]中的相同参数
RSLRS	使用文献[67]中的相同参数
E3WD	使用文献[68]中的相同参数
CART	使用 Python 中 Scikit-learn 库中的默认参数.
SVM	使用 Python 中 Scikit-learn 库中的默认参数.
NBC	使用 Python 中 Scikit-learn 库中的默认参数.
LR	使用 L2 正则化.
RF	使用十折交叉验证确定 *n_estimators* 的取值.
kNN	通过十折交叉验证确定 k 的取值.

从表 11～表 14 可以看出在大部分数据集上,基于本文方法构建的分类器相比其他 15 个分类器在性能上有着明显优势;从平均性能来看基于本文方法构建的分类器在 4 项通用的分类指标下均高于其他 15 个分类器,如图 8 所示;同时 STD 在 4 项指标下均低于其他 15 个分类器.实验结果表明,本文方法在面向分类任务在保证平均性能高于其他分类器的前提下,同时具有较强的稳定性.

特别地,与机器学习分类器相比,本文方法面向分类任务的分类性能和时间复杂度的对比分析如下:本文方法面向分类任务的时间复杂度为 $O(n\log n)$.CART 的时间复杂度为 $O(n\log n)$.SVM 的时间复杂度最好情况为 $O(n^2)$,最坏情况为 $O(n^3)$.NBC 的时间复杂度为 $O(nfc)$.LR 的时间复杂度为 $O(nf)$.RF 的时间复杂度为 $O(nfk\log n)$.kNN 的时间复杂度为 $O(nf\log n)$.其中,n 为对象数目,f 为属性数目,c 为决策类数目,k 为树的数目.由此可见:

表 11 *Accuracy* 对比

编号	本文方法	粒计算分类器									机器学习分类器					
		NCR	NCRDPP	TNCR	GBNRS	GB-kNN	FNC-TC	FNC-EC	RSLRS	E3WD	CART	SVM	NBC	LR	RF	kNN
1	1.0000	0.9533	0.9600	0.9533	0.9923	0.9452	0.9600	0.9923	0.9825	0.9533	0.9533	0.9600	0.9519	0.8195	0.9395	0.9519
2	0.8889	0.7037	0.7593	0.7667	0.8405	0.7741	0.8252	0.8370	0.8233	0.7334	0.7481	0.8370	0.8365	0.8407	0.7481	0.7981
3	0.9582	0.7973	0.8345	0.8420	0.9023	0.8332	0.8420	0.8725	0.8545	0.8599	0.7980	0.8394	0.7557	0.7740	0.7831	0.8420
4	1.0000	0.6275	0.9494	0.9595	0.9538	0.9457	0.9494	0.9595	0.9233	0.9494	0.9378	0.9538	0.9551	0.9624	0.9436	0.9256
5	1.0000	0.9993	0.9941	0.9577	0.9881	0.9770	0.9825	0.9954	0.9567	0.9851	0.9881	0.9881	0.8456	0.9666	0.9718	0.9545
6	0.9289	0.8703	0.9034	0.8929	0.8834	0.8663	0.7903	0.8703	0.8663	0.9025	0.8713	0.8523	0.7903	0.7859	0.8518	0.7903
7	0.9116	0.6681	0.7328	0.7652	0.8583	0.7841	0.8823	0.9116	0.8496	0.7967	0.7501	0.7638	0.7583	0.7735	0.7031	0.8703
8	0.7878	0.4883	0.6378	0.5833	0.7568	0.7600	0.7756	0.7722	0.7600	0.7098	0.7826	0.7665	0.7648	0.7691	0.7486	0.7645
9	0.8336	0.7796	0.8110	0.9604	0.8678	0.8098	0.8246	0.8336	0.8678	0.9632	0.7516	0.8696	0.8098	0.8678	0.7220	0.7796
10	0.9035	0.7973	0.9425	0.9460	0.8245	0.9450	0.8856	0.8945	0.9235	0.9520	0.8645	0.8915	0.8422	0.9020	0.8325	0.8856
11	0.9340	0.8394	0.9290	0.9460	0.8990	0.9460	0.8992	0.9225	0.8678	0.9460	0.9270	0.9510	0.8314	0.7723	0.9190	0.9412
12	0.9562	0.7481	0.9430	0.9460	0.9036	0.9460	0.9420	0.9520	0.9445	0.9036	0.9205	0.9555	0.7037	0.9520	0.9040	0.9523
13	0.8571	0.8235	0.8345	0.8245	0.8415	0.8312	0.8312	0.8499	0.8235	0.8345	0.8245	0.8235	0.8332	0.8199	0.8235	0.8401
14	0.9821	0.9547	0.9645	0.9512	0.9688	0.9322	0.8412	0.9820	0.9688	0.9320	0.9688	0.9549	0.9512	0.9323	0.9478	0.9349
15	0.9457	0.8885	0.8912	0.8841	0.9404	0.9178	0.9320	0.9323	0.9231	0.9149	0.8912	0.8685	0.8841	0.8456	0.8749	0.8812
16	0.8525	0.7745	0.7814	0.7847	0.8512	0.7918	0.7923	0.8456	0.7814	0.8125	0.7847	0.7495	0.7732	0.7745	0.7745	0.7925
17	0.8023	0.7136	0.7396	0.7136	0.7985	0.7547	0.7748	0.7799	0.7136	0.7014	0.7136	0.7356	0.7014	0.7136	0.7248	0.7410
18	0.8323	0.7211	0.7414	0.7323	0.8299	0.8007	0.7823	0.8032	0.8173	0.8007	0.7414	0.7211	0.7323	0.7189	0.7689	0.7686
平均值	0.9097	0.7860	0.8527	0.8561	0.8834	0.8645	0.8618	0.8892	0.8693	0.8680	0.8454	0.8601	0.8178	0.8328	0.8323	0.8563
STD	0.0671	0.1224	0.1004	0.1079	0.0643	0.0764	0.0672	0.0691	0.0731	0.0880	0.0873	0.0850	0.0775	0.0792	0.0869	0.0728

表 12 *Recall* 对比

编号	本文方法	粒计算分类器									机器学习分类器					
		NCR	NCRDPP	TNCR	GBNRS	GB-kNN	FNC-TC	FNC-EC	RSLRS	E3WD	CART	SVM	NBC	LR	RF	kNN
1	1.0000	0.9194	0.9600	0.9533	0.9600	0.9450	0.9322	0.9945	0.9235	0.9123	0.9533	0.9600	0.9573	0.8195	0.9395	0.9519
2	0.8888	0.7033	0.7644	0.7732	0.8888	0.7750	0.8252	0.8370	0.9567	0.7334	0.7416	0.8357	0.8354	0.8397	0.7407	0.7863
3	0.8829	0.6633	0.8460	0.8420	0.7633	0.7345	0.8011	0.8725	0.7533	0.8599	0.6481	0.7158	0.5447	0.5882	0.6011	0.8420
4	1.0000	0.5896	0.9409	0.9530	0.9659	0.9450	0.9494	0.9595	0.9494	0.8420	0.9263	0.9528	0.9590	0.9589	0.9344	0.9256
5	1.0000	0.9993	0.9940	0.9564	0.9893	0.9769	0.9825	0.9954	0.9567	0.9851	0.9882	0.9892	0.8413	0.9656	0.9712	0.9545
6	0.9478	0.8145	0.8558	0.8655	0.9429	0.7982	0.7903	0.8703	0.9123	0.7935	0.8294	0.8294	0.8576	0.8449	0.9256	0.7903
7	0.9120	0.5999	0.7328	0.7652	0.7353	0.6491	0.8823	0.9116	0.8496	0.8135	0.7353	0.5619	0.5744	0.5667	0.5786	0.8706
8	0.7911	0.5116	0.5359	0.3333	0.7155	0.7393	0.5532	0.7722	0.7158	0.7098	0.7887	0.5779	0.6003	0.5767	0.7155	0.7645
9	0.8321	0.7794	0.8109	0.8606	0.8697	0.8094	0.8246	0.8336	0.8606	0.8532	0.7519	0.8695	0.8092	0.8697	0.7220	0.7796
10	0.5818	0.4492	0.5178	0.5000	0.5818	0.4995	0.4032	0.5189	0.4235	0.5178	0.5155	0.4984	0.5421	0.5026	0.5250	0.5736
11	0.6627	0.6497	0.5113	0.5000	0.6610	0.5000	0.4832	0.5323	0.4928	0.5323	0.6479	0.6750	0.6589	0.5000	0.4929	0.5323
12	0.5886	0.5246	0.5206	0.5000	0.5523	0.5000	0.4215	0.4578	0.5467	0.5631	0.5159	0.4949	0.5631	0.5064	0.5190	0.5033
13	0.8612	0.8235	0.8345	0.8245	0.8415	0.8312	0.8312	0.8499	0.8235	0.8345	0.8245	0.8235	0.8332	0.8199	0.8235	0.8401
14	0.9699	0.9547	0.9645	0.9512	0.9688	0.9322	0.8412	0.9623	0.9156	0.9323	0.9688	0.9549	0.9512	0.9323	0.9478	0.9349
15	0.9457	0.8885	0.8912	0.8841	0.9404	0.9178	0.9320	0.9323	0.8912	0.9320	0.8912	0.8685	0.8841	0.8456	0.8213	0.8812
16	0.8400	0.7745	0.7814	0.7847	0.8512	0.7918	0.7923	0.8456	0.7814	0.7625	0.7847	0.7495	0.7323	0.7745	0.7745	0.7925
17	0.8532	0.7136	0.7396	0.7136	0.7985	0.7547	0.7748	0.7799	0.8512	0.7914	0.7136	0.7348	0.7014	0.7299	0.7248	0.7410
18	0.8832	0.7211	0.7414	0.7323	0.8299	0.8007	0.7823	0.8032	0.8173	0.7805	0.7345	0.7211	0.7323	0.7189	0.7689	0.7686
平均值	0.8578	0.7267	0.7746	0.7607	0.8253	0.7722	0.7668	0.8183	0.8012	0.7861	0.7755	0.7674	0.7543	0.7422	0.7515	0.7907
STD	0.1261	0.1530	0.1559	0.1804	0.1302	0.1483	0.1739	0.1560	0.1564	0.1312	0.1359	0.1541	0.1428	0.1576	0.1536	0.1316

表 13　*Precision* 对比

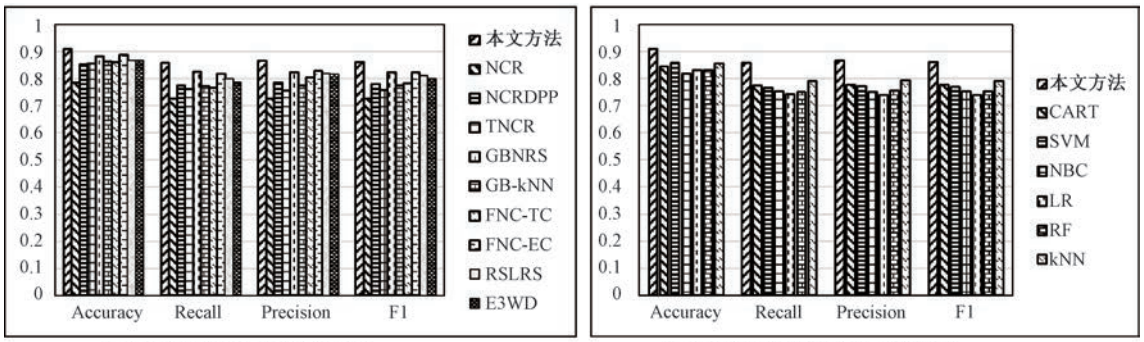
编号	本文方法	粒计算分类器									机器学习分类器					
		NCR	NCRDPP	TNCR	GBNRS	GB-kNN	FNC-TC	FNC-EC	RSLRS	E3WD	CART	SVM	NBC	LR	RF	kNN
1	1.0000	0.9171	0.9600	0.9533	0.9395	0.9571	0.9789	0.9989	0.9589	0.9823	0.9533	0.9600	0.9478	0.8195	0.9395	0.9519
2	0.8877	0.7222	0.7521	0.7638	0.8326	0.7822	0.8252	0.8812	0.8467	0.8543	0.7368	0.8256	0.8332	0.8326	0.7378	0.7981
3	0.8735	0.6743	0.8369	0.8420	0.7990	0.7517	0.8420	0.8725	0.7933	0.8520	0.6626	0.7221	0.5640	0.6051	0.6228	0.8420
4	1.0000	0.5907	0.9494	0.9587	0.9629	0.9491	0.9123	0.9595	0.9429	0.9320	0.9359	0.9498	0.9442	0.9601	0.9430	0.9256
5	1.0000	0.9992	0.9940	0.9584	0.9725	0.9792	0.9825	0.9954	0.9525	0.9854	0.9880	0.9873	0.8470	0.9674	0.9725	0.9545
6	0.9354	0.8234	0.9285	0.9363	0.8686	0.8585	0.7903	0.8703	0.9263	0.9332	0.8856	0.8971	0.7228	0.8296	0.8217	0.7903
7	0.9050	0.6037	0.7328	0.7652	0.8829	0.7077	0.8823	0.9116	0.8496	0.9035	0.7401	0.5770	0.6385	0.5496	0.5915	0.8703
8	0.7973	0.3631	0.5522	0.1961	0.7798	0.7212	0.7756	0.7722	0.7768	0.7798	0.7893	0.5792	0.6475	0.5457	0.7108	0.7645
9	0.8569	0.7786	0.8109	0.8610	0.8676	0.8108	0.8246	0.8336	0.8708	0.8346	0.7517	0.8694	0.8415	0.8676	0.7219	0.7796
10	0.5788	0.4710	0.5503	0.5430	0.5223	0.4730	0.5612	0.5389	0.4935	0.5178	0.5166	0.4621	0.5542	0.4958	0.5219	0.5032
11	0.7134	0.6225	0.5279	0.4730	0.5009	0.4730	0.5512	0.5596	0.5028	0.5323	0.6161	0.7358	0.6088	0.4813	0.5040	0.5523
12	0.6847	0.5877	0.5905	0.4730	0.6726	0.4730	0.5923	0.5832	0.6467	0.5631	0.5216	0.4792	0.5289	0.5295	0.6618	0.6012
13	0.8612	0.8235	0.8345	0.8245	0.8415	0.8312	0.8312	0.8499	0.8435	0.8345	0.8245	0.8235	0.8332	0.8199	0.8235	0.8401
14	0.9699	0.9547	0.9645	0.9512	0.9688	0.9322	0.8412	0.9523	0.9556	0.9012	0.9688	0.9549	0.9512	0.9323	0.9478	0.9349
15	0.9457	0.8885	0.8912	0.8841	0.9404	0.9178	0.9320	0.9323	0.8912	0.9320	0.8912	0.8685	0.8841	0.8456	0.8213	0.8812
16	0.8392	0.7745	0.7814	0.7847	0.8512	0.7918	0.7923	0.8456	0.8554	0.8025	0.7847	0.7495	0.7323	0.7745	0.7745	0.7925
17	0.8532	0.7621	0.7396	0.7136	0.7985	0.7547	0.7748	0.7799	0.8512	0.7714	0.7136	0.7348	0.7014	0.7299	0.7248	0.7410
18	0.8832	0.7211	0.7414	0.7323	0.8299	0.8007	0.7823	0.8032	0.8173	0.7805	0.7345	0.7211	0.7323	0.7189	0.7689	0.7686
平均值	0.8658	0.7266	0.7855	0.7563	0.8240	0.7758	0.8040	0.8300	0.8208	0.8162	0.7786	0.7721	0.7507	0.7392	0.7561	0.7940
STD	0.1112	0.1629	0.1476	0.2046	0.1332	0.1572	0.1228	0.1368	0.1361	0.1401	0.1387	0.1585	0.1365	0.1596	0.1379	0.1270

表 14　*F*1 对比

编号	本文方法	粒计算分类器									机器学习分类器					
		NCR	NCRDPP	TNCR	GBNRS	GB-kNN	FNC-TC	FNC-EC	RSLRS	E3WD	CART	SVM	NBC	LR	RF	kNN
1	1.0000	0.9182	0.9600	0.9533	0.9496	0.9510	0.9550	0.9967	0.9409	0.9460	0.9533	0.9600	0.9525	0.8195	0.9395	0.9519
2	0.8882	0.7126	0.7582	0.7685	0.8598	0.7786	0.8252	0.8585	0.8983	0.7892	0.7392	0.8306	0.8343	0.8361	0.7392	0.7922
3	0.8782	0.6688	0.8414	0.8420	0.7807	0.7430	0.8210	0.8725	0.7728	0.8559	0.6553	0.7189	0.5542	0.5965	0.6118	0.8420
4	1.0000	0.5901	0.9451	0.9558	0.9644	0.9470	0.9305	0.9595	0.9461	0.8847	0.9311	0.9513	0.9515	0.9595	0.9387	0.9256
5	1.0000	0.9992	0.9940	0.9574	0.9808	0.9780	0.9825	0.9954	0.9546	0.9852	0.9881	0.9882	0.8441	0.9665	0.9718	0.9545
6	0.9416	0.8189	0.8907	0.8995	0.9042	0.8273	0.7903	0.8703	0.9192	0.8577	0.8566	0.8619	0.7845	0.8372	0.8706	0.7903
7	0.9085	0.6018	0.7328	0.7652	0.8024	0.6771	0.8823	0.9116	0.8496	0.8561	0.7377	0.5693	0.6048	0.5580	0.5850	0.8704
8	0.7942	0.4247	0.5439	0.2469	0.7463	0.7301	0.6458	0.7722	0.7451	0.7432	0.7890	0.5785	0.6230	0.5608	0.7131	0.7645
9	0.8443	0.7790	0.8109	0.8608	0.8686	0.8101	0.8246	0.8336	0.8657	0.8438	0.7518	0.8694	0.8250	0.8686	0.7219	0.7796
10	0.5803	0.4598	0.5336	0.5206	0.5504	0.4859	0.4693	0.5287	0.4558	0.5178	0.5160	0.4796	0.5481	0.4992	0.5234	0.5361
11	0.6871	0.6358	0.5195	0.4861	0.5699	0.4861	0.5150	0.5456	0.4977	0.5323	0.6316	0.7041	0.6329	0.4905	0.4984	0.5421
12	0.6330	0.5544	0.5534	0.4861	0.6065	0.4861	0.4925	0.5129	0.5925	0.5631	0.5187	0.4869	0.5455	0.5177	0.5818	0.5479
13	0.8612	0.8235	0.8345	0.8245	0.8415	0.8312	0.8312	0.8499	0.8334	0.8345	0.8245	0.8235	0.8332	0.8199	0.8235	0.8401
14	0.9699	0.9547	0.9645	0.9512	0.9688	0.9322	0.8412	0.9573	0.9352	0.9165	0.9688	0.9549	0.9512	0.9323	0.9478	0.9349
15	0.9457	0.8885	0.8912	0.8841	0.9404	0.9178	0.9320	0.9323	0.8912	0.9320	0.8912	0.8685	0.8841	0.8456	0.8213	0.8812
16	0.8396	0.7745	0.7814	0.7847	0.8512	0.7918	0.7923	0.8456	0.8167	0.7820	0.7847	0.7495	0.7323	0.7745	0.7745	0.7925
17	0.8532	0.7371	0.7396	0.7136	0.7985	0.7547	0.7748	0.7799	0.8512	0.7813	0.7136	0.7348	0.7014	0.7299	0.7248	0.7410
18	0.8832	0.7211	0.7414	0.7323	0.8299	0.8007	0.7823	0.8032	0.8173	0.7805	0.7345	0.7211	0.7323	0.7189	0.7689	0.7686
平均值	0.8616	0.7257	0.7798	0.7574	0.8230	0.7738	0.7827	0.8237	0.9102	0.8001	0.7770	0.7695	0.7519	0.7406	0.7531	0.7920
STD	0.1187	0.1582	0.1515	0.1945	0.1293	0.1525	0.1505	0.1468	0.1456	0.1330	0.1371	0.1559	0.1384	0.1586	0.1449	0.1293

(1) 相比 CART、SVM、RF 和 kNN 分类器,本文方法面向分类任务的时间复杂度较低;

(2) 相比 NBC 和 LR 分类器,本文方法面向分类任务的时间复杂度不具有优势,但是根据表 11～表 14,相比 NBC 分类器,本文方法面向分类任务在 *Accuracy* 指标下提升了 11.12%,在 *Recall* 指标下提升了 13.72%,在 *Precision* 指标下提升了 15.33%,在 $F1$ 指标下提升了 14.67%;相比 LR 分类器,本文方法面向分类任务的时间复杂度不具有优势,但是根据表 11～表 14,相比 NBC 分类器,本文方法面向分类任务在 *Accuracy* 指标下提升了 9.23%,在 *Recall* 指标下提升了 15.87%,在 *Precision* 指标下提升了 17.13%,在 $F1$ 指标下提升了 16.34%.



(a) 本文方法与粒计算分类器对比 (b) 本文方法与机器学习分类器对比

图 8 在四个指标下模型性能的均值对比

4.4 面向分类任务的显著性评估

为了验证本文方法用于分类任务相比其他分类器在分类性能方面的显著性，本节将在 $\alpha=0.1$ 的显著水平上进行 *Friedman* 检验. 表 15 展示了 16 个分类器在四个分类指标下的平均序值和 *Friedman* 检验的结果 τ_F .

表 15 *F1* 对比

指标	本文方法	粒计算分类器									机器学习分类器						τ_F
		NCR	NCRDPP	TNCR	GBNRS	GB-*k*NN	FNC-TC	FNC-EC	RSLRS	E3WD	CART	SVM	NBC	LR	RF	*k*NN	
Accuracy	1.81	12.11	8.00	7.94	5.39	8.33	7.94	4.28	8.39	8.11	10.00	8.50	12.06	11.06	12.67	9.33	9.86
Recall	1.11	11.39	8.00	9.61	3.50	9.61	9.89	4.56	8.50	8.78	9.44	10.00	9.78	11.50	11.28	9.00	10.40
Precision	1.61	10.67	8.17	9.56	5.83	10.22	7.78	4.61	6.39	6.67	9.72	9.83	11.06	12.22	11.94	9.67	9.62
*F*1	1.56	11.44	8.00	9.72	4.06	10.22	9.11	4.39	7.28	8.06	9.61	9.61	10.56	12.00	11.56	8.33	10.36

由于算法数目 $M=16$，数据集数目 $N=18$，由表 15 可知，$\tau_F \gg 1.770$，这表明 16 个分类器之间存在显著性差异. 因此，需要进行 *Nemenyi* 事后检验来确定任何两种方法之间是否存在显著性差异. *Nemenyi* 检验的中，获得的临界距离(Critical distance，CD)如下：

$$CD=q_\alpha\sqrt{\frac{N(N+1)}{6N}}.$$

其中，$q_{\alpha=0.01}=2.2299$. 当两种比较方法之间的距离超过 $CD=2.7562$ 时，分类性能存在显著差异.

15 种分类器排序的 CD 图如图 9 所示，其中分类器的排序越小，性能越好. 如图 9 所示，本文方法面向分类任务在所有指标上排名均为第一，因此据统计分析，本文方法面向分类任务在大多数情况下优于其他比较方法；同时，在 $CD=2.7562$ 的显著性条件下，本文方法面向分类任务在 *Accuracy*、*Pre-*

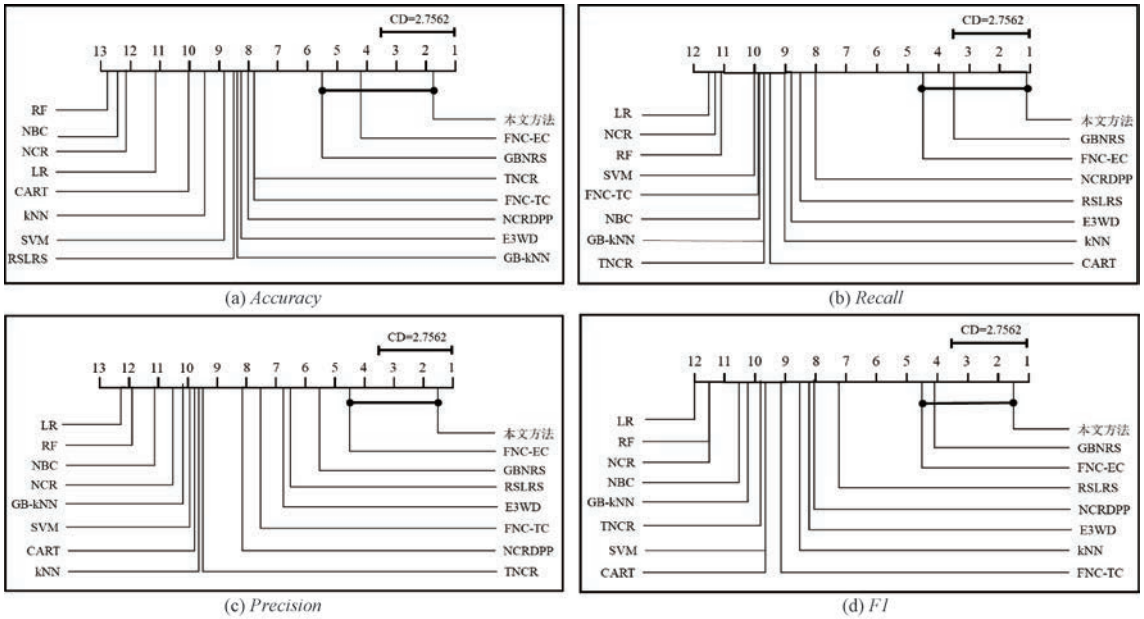


图 9 Nemenyi 检验在四个分类指标下的 CD 图

cision、*F*1 三个分类指标下与 12 个分类器具有显著性差异,在 *Recall* 分类指标下与 11 个分类器具有显著性差异.因此验证了本文方法面向分类任务的有效性和显著性.

综上所述,本文提出的基于 2 近邻模糊关系的多粒度空间快速构建方法不仅提升了模糊商空间模型构建多粒度空间的效率,同时能够构建出与现有模糊商空间模型完全一致的多粒度空间,保证了与现有模糊商空间模型相同的分类能力;相比现有常用的粒计算分类器和机器学习分类器具有较高的分类能力、较强的稳定性和一定的统计显著性.

5　结　论

作为模拟人类认知模式的典型 GrC 模型,模糊商空间理论基于大数据构建多粒度空间,已成功用于知识发现.但由于模糊商空间理论的步骤 1 模糊相似关系的构建需要耗费大量时间,且模糊相似关系中的信息存在大量冗余.这不仅是一种无效的时间损耗,同时影响后续步骤的效率.因此,本文首先提出了 k 近邻模糊关系,并分析其自反性、对称性和传递性;针对 k 的取值,通过分析模糊相似关系有效值的提取算法,从理论上证明了 2 近邻模糊关系就已包含模糊相似关系全部的有效值,并等价于模糊相似关系对应的模糊等价关系;然后,基于 2 近邻模糊关系,提出了论域的最近邻序列和次近邻序列,提出了论域的最近邻序列和次近邻序列,基于相互最近邻数设计最近邻阶段的有效位置数计算方法,基于统计学的频数概念设计次近邻阶段的有效位置数计算方法,并在此基础上提出了两阶段的有效值有效位置提取算法;基于上述理论分析,提出了多粒度空间的快速构建方法;最后,通过在 9 个 UCI 数据集、3 个 UKB 数据集、3 个图像数据集和 3 个文本数据集上的对比实验,验证了该算法构建多粒度空间的效率,通过同现有分类器进行对比分析,说明了本文方法面向包括辅助医疗诊断、图像分类和文本分类等场景下分类任务的有效性、稳定性和显著性.综上所述,本文提出了一种基于 2 近邻模糊关系的多粒度空间快速构建方法,在保证分类效果的前提下大幅降低了时间复杂度.然而,相似性度量的计算方式决定了本文方法在增量场景下无法直接应用,在未来研究中,我们将研究基于模糊商空间理论的增量学习机制.

参　考　文　献

[1] Wang Guo-Yin, Fu Shun, Yang Jie, Guo Yi-Ke. A review of research on multi-granularity cognition based intelligent computing. Chinese Journal of Computers, 2022, 45(6): 1161-1175. (in Chinese)
(王国胤, 傅顺, 杨洁, 郭毅可. 基于多粒度认知的智能计算研究. 计算机学报, 2022, 45(6): 1161-1175)

[2] Pedrycz W, Homenda W. Building the fundamentals of granular computing: A principle of justifiable granularity. Applied Soft Computing, 2013, 13(10): 4209-4218

[3] Liang Ji-Ye, Qian Yu-Hua, Li De-Yu, Hu Qing-Hua. Theory and method of granular computing for big data mining. Scientia Sinica (Informationis), 2015, 45(11): 1355-1369. (in Chinese)
(梁吉业, 钱宇华, 李德玉, 胡清华. 大数据挖掘的粒计算理论与方法. 中国科学:信息科学, 2015, 45(11): 1355-1369)

[4] Ciucci D, Yao Y Y. Synergy of granular computing, shadowed sets, and three-way decisions. Information Sciences, 2020, 508: 422-425

[5] Yao Y Y. Three-way granular computing, rough sets, and formal concept analysis. International Journal of Approximate Reasoning, 2020, 116: 106-125

[6] Zadeh L A. Fuzzy sets and information granularity. Advances in Fuzzy Set Theory and Applications, 1979, 11: 3-18

[7] Lin T Y. Granular computing on binary relations: Rough set representations and belief functions. Rough Sets in Knowledge Discovery, 1998, 1: 121-140

[8] Han J, Micheline K. Data mining: Concepts and techniques. Data Mining Concepts Models Methods and Algorithms Second Edition, 2006, 5(4): 1-18

[9] Zhou J, Pedrycz W, Gao C, Lai Z, Yue X. Principles for constructing three-way approximations of fuzzy sets: A comparative evaluation based on unsupervised learning. Fuzzy Sets and Systems, 2020, 413, 74-98

[10] Xu K J, Pedrycz W, Li Z W, et al. Constructing a virtual space for enhancing the classification performance of fuzzy clustering. IEEE Transactions on Fuzzy Systems, 2018, 27(9): 1779-1792

[11] Pedrycz W. Granular Computing: Analysis and Design of Intelligent Systems. Boca Raton, USA: CRC Press, 2013

[12] Guo L, Zhan J M, Xu Z S, et al. A three-way consensus model with regret theory under the framework of probabilistic linguistic term sets. Information Sciences, 2023, 632: 144-163

[13] Zhu J X, Ma X A, Herrera-Viedma E, Zhan J M. A three-way consensus model with regret theory under the framework of probabilistic linguistic term sets. Information Fusion, 2023, 95: 250-274

[14] Zhang B, Zhang L. Theory and Applications of Problem Solving. Amsterdam, the Netherland: Elsevier Science Inc., 1992

[15] Zhang B, Zhang L. Theory of fuzzy quotient space (methods of fuzzy granular computing). Journal of Software, 2003, 14(4): 770-776

[16] Zhang B, Zhang L. Fuzzy reasoning model under quotient space structure. Information Sciences, 2005, 173(4): 353-364

[17] Lin G, Liang J Y, Qian Y H, Li J. A fuzzy multi-granulation decision-theoretic approach to multi-source fuzzy information systems. Knowledge-Based Systems, 2016, 91: 102-113

[18] Li Jin Hai, Mi Yun Long, Liu Wen Qi. Incremental cognition of concepts: Theories and methods. Chinese Journal of Computers, 2019, 42(10): 2233-2250. (in Chinese)
(李金海, 米允龙, 刘文奇. 概念的渐进式认知理论与方法. 计算机学报, 2019, 42(10): 2233-2250)

[19] Zhang B, Zhang L. The structure analysis of fuzzy sets. International Journal of Approximate Reasoning, 2005, 40(2): 92-108

[20] Miao D Q, Fan S D. The calculation of knowledge granulation and its application. System Engineering Theory and Practice, 2022, 22(1): 48-56

[21] Qian Y H, Liang J Y, Yao Y Y, et al. Mgrs: A multi-granulation rough set. Information Sciences, 2010, 180(6): 949-970

[22] Qian Y H, Liang J Y, Dang C Y. Incomplete multi-granulation rough set. IEEE Press, 2010, 40, 420-431

[23] Lin G P, Liang J Y, Qian Y H. An information fusion approach by combining multi-granulation rough sets and evidence theory. Information Sciences, 2015, 314: 184-199

[24] Zhao S, Zhang L, Xu X, Zhang Y. Hierarchical description of uncertain information. Information Sciences, 2014, 268: 133-146

[25] Huang B, Guo C X, Zhuang Y L, Li H X, Zhou X Z. Intuitionistic fuzzy multi-granulation rough sets. Information Sciences, 2014, 277: 299-320

[26] Yang X, Qi Y, Song X, Yang J. Test cost sensitive multigranulation rough set: Model and minimal cost selection. Information Sciences, 2013, 250: 184-199

[27] Wu X Z, Leung Y. Theory and applications of granular labelled partitions in multi-scale decision tables. Information Sciences, 2011, 181(18): 3878-3897

[28] Li F, Hu B Q. A new approach of optimal scale selection to Bmulti-scale decision tables. Information Sciences, 2016, 381: 193-208

[29] Zhang Q H, Cheng Y L, Zhao F, Wang G Y, Xia S Y. Optimal scale combination selection integrating three-way decision with hasse diagram. IEEE Transactions on Neural Networks and Learning Systems, 2022, 3(8): 3675-3689

[30] Zhang Qing Hua, Zhi Xue Chao, Wang Guo Yin, et al. Multi-granularity ensemble classification algorithm based on attribute representation. Chinese Journal of Computers, 2022, 45(8): 1712-1729. (in Chinese)
(张清华, 支学超, 王国胤等. 基于属性代表的多粒度集成分类算法. 计算机学报, 2022, 45(8): 1712-1729)

[31] Xia S Y, Zhang H, Li W, et al. GBNRS: A novel rough set algorithm for fast adaptive attribute reduction in classification. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(3): 1231-1242

[32] Wu Cheng Ying, Zhang Qing Hua, Zhao Fan, et al. Hyperinterval granulation approach based on the density peaks clustering for classification. Chinese Journal of Computers, 2023, 46(8): 1620-1635. (in Chinese)
(吴成英, 张清华, 赵凡等. 基于密度峰值聚类的超区间粒化方法及其分类模型. 计算机学报, 2023, 46(8): 1620-1635)

[33] Zhang Q H, Zhao F, Cheng Y L, et al. Effective value analysis of fuzzy similarity relation in HQSS for efficient granulation. IEEE Transactions on Neural Networks and Learning Systems, 2023, DOI:10.1109/TNNLS.2023.3265310

[34] Liang J Y, Wang J, Qian Y H. A new measure of uncertainty based on knowledge granulation for rough sets. Information Sciences, 2009, 9(4): 458-470

[35] Liang J Y, Qian Y H, Dang C Y. Knowledge structure, knowledge granulation and knowledge distance in a knowledge base. International Journal of Approximate Reasoning, 2009, 50(1): 174-188

[36] Yao Y Y, Zhao L Q. A measurement theory view on the granularity of partitions. Information Sciences, 2012, 213(5): 1-13

[37] Zhang Qing-Hua, Wang Guo-Yin, Liu Xian-Quan. Hierarchical structure analysis of fuzzy quotient space. Pattern Recognition and Artificial Intelligence, 2008, 21(5): 627-634. (in Chinese)
(张清华, 王国胤, 刘显全. 分层递阶的模糊商空间结构分析. 模式识别与人工智能, 2008, 21(5): 627-634)

[38] Zhang Q H, Wang G Y. The uncertainty measure of hierarchical quotient space structure. Mathematical Problems in Engineering, 2011: 505-515

[39] Zhang Q H, Yang S H, Wang G Y. Measuring uncertainty of probabilistic rough set model from its three regions. IEEE Transactions on Systems Man, and Cybernetics: Systems, 2016, 47(12): 3299-3309

[40] Zhang Q H, Xu K, Wang G Y. Fuzzy equivalence relation and its multi-granulation spaces. Information Sciences, 2016, 346: 44-57

[41] Yang J, Wang G Y, Zhang Q H. Knowledge distance measure in multi-granulation spaces of fuzzy equivalence relations. Information Sciences, 2018, 448: 18-35

[42] Zhao F, Zhang Q H, Wu C Y, et al. A neighborhood covering classifier based on optimal granularity of fuzzy quotient space. IEEE Transactions on Fuzzy Systems, 2023, 31(10): 3567-3581

[43] Pedrycz A, Reformat M. Hierarchical fcm in a stepwise discovery of structure in data. Soft Computing, 2006, 10(3): 244-256

[44] Tsekouras G, Sarimveis H, Kavakli E, et al. A hierarchical fuzzy-clustering approach to fuzzy modeling. Fuzzy Sets and Systems, 2005, 150(2): 245-266

[45] Zhang Qing-Hua. A hierarchical fuzzy decision-making method. Microelectronics and Computer, 2009, 26(2): 124-127, 132. (in Chinese)
(张清华. 一种分层递阶的模糊决策方法. 微电子学与计算机, 2009, 26(2): 124-127,132)

[46] Tang X Q, Ping Z. Hierarchical clustering problems and analysis of fuzzy proximity relation on granular space. IEEE Transactions on Fuzzy Systems, 2013, 21(5): 814-824

[47] Cui G Z, Yu J P, Wang Q G. Finite-time adaptive fuzzy control for mimo nonlinear systems with input saturation via improved command-filtered backstepping. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(2): 980-989

[48] Pei D, Rui Y. Hierarchical structure and applications of fuzzy logical systems. International Journal of Approximate Reasoning, 2013, 54(9): 1483-1495

[49] Chang C W, Tao C W. A simplified implementation of hierarchical fuzzy systems. Soft Computing, 2019, 23(12): 4471-4481

[50] Wang L X. Hierarchical fuzzy opinion networks: Topcdown for social organizations and bottomcup for election. IEEE Transactions on Fuzzy Systems, 2020, 28(7): 1265-1275

[51] Liu Ren Jin, Huang Xian Wu. The granular theorem of quotient space in image segmentation. Chinese Journal of Computers, 2005, 28(10): 1680-1685. (in Chinese)
(刘仁金,黄贤武.图像分割的商空间粒度原理. 计算机学报, 2005, 28(10): 1680-1685)

[52] Tan Y H, Chan C S. Phrase-based image caption generator with hierarchical lstm network. Neurocomputing, 2019, 333:86-100

[53] Page G L, Rodriguez-Varez M X, Lee D. Bayesian hierarchical modelling of growth curve derivatives via sequences of quotient differences. Journal of the Royal Statistical Society, 2020, 69: 459-481

[54] Erick D, Yu W. Data-driven fuzzy modeling using restricted boltzmann machines and probability theory. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 50(7): 2316-2326

[55] Yan J J, Yang J H, Li X J. Adaptive fault-tolerant compensation control for t-s fuzzy systems with mismatched parameter uncertainties. IEEE Transactions on Systems Man, and Cybernetics: Systems, 2020, 50(9): 3412-3423

[56] Zhang Q H, Chen Y H, Yang J, et al. Fuzzy entropy: A more comprehensible perspective for interval shadowed sets of fuzzy sets. IEEE Transactions on Fuzzy Systems, 2020, 28(11): 3008-3022

[57] Zhu Q S, Feng J, Huang J L. Natural neighbor: A self-adaptive neighborhood method without parameter k. Pattern Recognition Letters, 2016, 80: 30-36

[58] Roussopoulos N, Kelley S, Vincent F. Nearest neighbor queries//Proceedings ACM SIGMOD'95 International Conference on Management of Data. California: ACM, 1995: 71-79

[59] Brito M R, Chavez E L, Quiroz A J, et al. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. Statistics and Probability Letters, 1997, 35(1): 33-42

[60] Yong D, Hu Q H, Zhu P F, et al. Rule learning for classification based on neighborhood covering reduction. Information Sciences, 2011, 181(24): 5457-5467

[61] Yue X D, Xiao X, Chen Y F, et al. Robust neighborhood covering reduction with determinantal point process sampling. Knowledge-Based Systems, 2020, 188: 105063

[62] Yue X D, Chen Y F, Miao D Q. Tri-partition neighborhood covering reduction for robust classification. International Journal of Approximate Reasoning, 2017, 83: 371-384

[63] Xia S Y, Liu Y S, Xin D, et al. Granular ball computing classifiers for efficient, scalable and robust learning. Information Sciences, 2019, 483: 136-152

[64] Xia S Y, Dai X C, Wang G Y, et al. An efficient and adaptive granular-ball generation method in classification problem. 2022, doi:10.48550/arXiv.2201.04343

[65] Yue X D, Chen Y F, Miao D Q. Fuzzy neighborhood covering for three-way classification. Information Sciences, 2020, 507: 795-808

[66] Zhang Q H, Ai Z H, Zhang J Z, et al. A novel fast constructing neighborhood covering algorithm for efficient classification. Knowledge-Based Systems, 2021, 225: 107104

[67] Xia S Y, Bai X Y, Wang G Y, et al. An efficient and accurate rough set for feature selection, classification, and knowledge representation. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(8): 7724-7735

[68] Qian J, Wang D, Yu, Y, et al. E3WD: A three-way decision model based on ensemble learning. Information Sciences, 2024, 667: 120487

[69] Breiman L I, Friedman J H, Olshen R A, et al. Classification and regression trees. Biometrics, 1984, 40(3): 358-361

[70] Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20(3): 273-297

[71] Zhang H. The optimality of naive Bayes. Aa, 2004, 1(2): 3-9

[72] Howel D, Kleinbaum D G. Logistic Regression: A self learning text. Journal of the Royal Statistical Society Series D (The Statistician), 1995, 44(3): 410-411

[73] Breiman L. Random forests. Machine learning, 2001, 45: 5-32

[74] Bentley J L. Multidimensional binary search trees used for associative searching. Communications of the ACM, 1975, 18(9): 509-517

ZHAO Fan, Ph. D. candidate. Her research interests include uncertain information processing and measurement.

ZHANG Qing-Hua, Ph. D. , professor, Ph. D. supervisor. His research interests include granular computing, data mining and uncertain information processing.

WU Cheng-Ying, Ph. D. Her research interests include granular computing and supervised learning.

XIE Qin, Ph. D. candidate. Her research interests include knowledge discovery and granular computing.

WANG Guo-Yin, Ph. D. , professor, Ph. D. supervisor, CCF Fellow, Yangtze River Scholar. His main research interests include granular computing, data mining and neural network.

Background

Granular computing is a state-of-the-art methodology that simulates the multi-granularition cognitive pattern of the human brain to deal with complex problems. As a typical description of granular computing, fuzzy quotient space theory focuses on gradually granulating complex problems into the hierarchical multi-granularition spaces, thereby implementing hierarchical solution of the complex problems.

However, when dealing with massive high-dimensional data, the efficiency of constructing multi-granularition spaces through the fuzzy similarity relations in the existing fuzzy quotient space methods reduces significantly. On one hand, the fuzzy similarity relation is obtained by calculating the similarity among all objects, which is not conducive to processing large datasets; On the other hand, the fuzzy similarity relation contains a large amount of redundant information, which leads to a large number of redundant computation in the subsequent steps.

Therefore, based on the 2-nearest neighbor fuzzy relation, an efficient construction approach for constructing multi- granularition spaces is proposed, which greatly improves the efficiency on the premise of ensuring the performance when facing downstream classification tasks. First, based on the k-nearest neighbor algorithm, a-nearest neighbor fuzzy relation is proposed, and its key properties are discussed and proven. Second, for the multi-granularition spaces construction task, parameter analysis is performed on the-nearest neighbor fuzzy relation, theoretically proving that when is taken as 2, all effective information in the data space could be included. Then, the number of effective positions in the nearest neighbor and second nearest neighbor phases are defined. And the algorithm for extracting effective values and effective positions of the fuzzy similarity relation is proposed, improving the efficiency of constructing multi-granularition spaces. Finally, relevant experiments are conducted on 9 UCI datasets, 3 UKB datasets, 3 image datasets and 3 text datasets to validate the efficiency of multi-granularition spaces construction approach. By comparing and analyzing with the existing classifiers, the effectiveness, stability, and saliency of the proposed approach for classification tasks are demonstrated. In summary, a-nearest neighbor fuzzy relation that only contains sparse effective information is constructed. On the basis, an efficient construction approach for constructing multi-granularition spaces based on 2-nearest neighbor fuzzy relation is proposed, which greatly reduces time complexity while ensuring classification performance.

This work is supported by the National Natural Science Foundation of China (No. 62276038, No. 62221005), the Joint Fund of Chongqing Natural Science Foundation for Innovation and Development (No. CSTB2023NSCQ-LZX0164), the Chongqing Talent Program (No. CQYC20210202215).